Универзитет у Нишу

Електронски Факултет

Данијела Р. Алексић

РАЗВОЈ КОДЕРА ТАЛАСНОГ ОБЛИКА ЗА ПОТРЕБЕ НЕУРОНСКИХ МРЕЖА И ОБРАДУ СИГНАЛА

ДОКТОРСКА ДИСЕРТАЦИЈА

Ниш, 2022.

University of Niš

Faculty of Electronic Engineering

Danijela R. Aleksić

DEVELOPMENT OF WAVEFORM CODERS FOR NEURAL NETWORKS APPICATIONS AND SIGNAL PROCESSING

DOCTORAL DISSERTATION

Niš, 2022.

Захвалница

Non progredi est regredi!

За прогрес у многим сферама живота је потребно много мотивације, енергије и рада. Поред свега наведеног, за израду ове докторске тезе, је била неприкосновена подршка која ми је примарно пружена од мог ментора Проф. др Зорана Перића као и чланова његовог тима.

Такође бих се захвалила и компанији „Телеком Србија“, која ме је подржала у намери да своје пословно искуство проширим и овим академским.

На крају бих се посебно захвалила мојој породици, која је највише веровала у мене и подржавала ме у изради ове докторске тезе.

Аутор

Подаци о докторској дисертацији

Ментор: Др Зоран Перић, редовни професор, Универзитет у Нишу, Електронски факултет

Наслов: Развој кодера таласног облика за потребе неуронских мрежа и обраду сигнала

Резиме: Ова докторска дисертација има за циљ да дизајнира ниско-битне скаларне квантизере и анализира њихову примену у неуронским мрежама и обради сигнала. У овој тези разматрамо могућности и ограничења која почивају на квантизацији, као водећој техници за кодовање и компресију података. Посебно испитујемо неизбежан губитак тачности презентације сигнала и података услед квантизације у области обраде сигнала, као и код многих савремених решења које користе квантизацију. Као што је наведено у овој тези, постоје бројни квалитативни перформансни индикатори, који указују да одговарајућа параметризација квантизера може оптимизовати количине пренетих података у битима. Квантоване неуронске мреже представљају актуелну област истраживања, посебно значајну за уређаје са ограниченим ресурсима. Ослањајући се на мноштво закључака о скаларним квантизерима развијеним за обраду сигнала и узимајући у обзир предности скаларне квантизације, антиципирамо да ће студиозним сагледавањем статистичких карактеристика параметара неуронских мрежа, ова дисертација допринети циљу проналажења ефикасног решења компресије тежина неуронских мрежа употребом нових, добро пројектованих

	скаларних квантизера, примарно за квантизацију тежина неуронских мрежа у фази након тренинга.
Научна област:	Електротехничко и рачунарско инжењерство
Научна дисциплина:	Телекомуникације (Дигитална обрада и кодовање сигнала)
Кључне речи:	Скаларна квантизација, нискобитни квантизери, Лапласова функција густине вероватноће, однос сигнал-шум квантизације, неуронске мреже, тачност неуронских мрежа, квантоване неуронске мреже, пост-тренинг.
УДК:	621.391:004 (075.8)(076)
ЦЕРИФ класификација:	T 121
Тип лиценце креативне заједнице:	CC BY-NC-ND

Data on Doctoral Dissertation

Doctoral Supervisor:	Dr. Zoran Perić, Full Professor, University of Niš, Faculty of Electronic Engineering
Title of Doctoral Dissertation:	Development of waveform coders for neural networks applications and signal processing
Abstract:	This doctoral thesis aims to design low-bit scalar quantizers and analyze their application in Neural Networks (NNs) and signal processing. In this thesis, we consider the possibilities and limitations that rest on quantization, as a leading technique for data coding and compression. In particular, we examine the inevitable accuracy loss of signal and data presentation due to quantization in the signal processing area, as well as in many modern solutions, that use quantization. As stated in this thesis, there are a number of qualitative performance indicators, which indicate that appropriate quantizer parameterization can optimize the amount of data transmitted in bits. Quantized Neural Networks (QNNs) is a promising research area, especially important for resource constrained devices. Relying on a plethora of conclusions about scalar quantizers derived for signal processing tasks and taking into account the advantages of scalar quantization, we anticipate that by studying the statistical characteristics of neural network parameters, this thesis will contribute to determining an efficient weights compression solution utilizing new, well-designed scalar quantizers for post-training quantization.

Scientific Field:	Electrical and Computer Engineering
Scientific Discipline:	Telecommunication (Digital processing and signal coding)
Key Words:	Scalar Qunatization, Low-bit quantizer, Laplacian probability density function, SQNR, Neural Network, Accuracy of Neural Network, Quantized Neual Network, Post-Training.
UDC:	621.391:004 (075.8)(076)
CERIF Classification:	T 121
Creative Commons Licence Type:	CC BY-NC-ND

САДРЖАЈ:

1. УВОД

Дигитално кодовање је неминовност дигиталног доба у коме живимо. Окружени смо бројним дигиталним садржајима који су свеприсутни у свакодневном животу, као и у интеракцијама са машинама и другим људима. Дигитално кодовање информација је процес формирања компримованог дигиталног приказа информација са циљем што ефикаснијег преноса и чувања, тј. складиштења истог. Развој техника кодовања је мотивисан све већим обимом дигиталних садржаја, али и све захтевнијом латентношћу, ограничењима меморије и потребама за капацитетом транспортних комуникационих система [1]-[10]. Бројна апликативна решења као и корисничко искуство и потребе, додатно изискују развој нових алгоритама за кодовање и компресију. Сама дигитализација информација се реализује кроз три основне фазе: одмеравање, квантовање и кодовање [1]-[5]. Без умањења важности сваке од фаза, у дисертацији се посебна пажња посвећује поступку квантизације и пројектовању квантизера таласног облика за потребе компресије неуронских мрежа (NN) и обраде сигнала, уз постизање жељених нивоа компресије, али и тачности.

Кодовање се генерално односи на процес издвајања сирових, аналогних информација из сигнала и њиховог претварања у дигитално кодоване приказе, како би се олакшао приступ инваријантним атрибутима тог сигнала [2], [3], [9]. Декодовање сигнала, као реципрочан процес, има главни задатак да реконструише пријемни сигнал што је могуће ближе оригиналном сигналу, користећи екстраховане атрибуте. Добро дизајниран квантизер може ублажити потенцијални утицај променљивих статистичких карактеристика улазног сигнала на квалитет његове квантоване презентације, чиме доприноси повећаној робусности. Применом неког од закона компресије, улазни квантовани сигнал се приказује редукованим бројем битова, одржавајући или пожељно, незнатно деградирајући тачност и квалитета сигнала [3] - [6], [8].

Квантизер се може представити редном везом енкодера и декодера [2], а функционално подразумева пресликавање амплитуда сигнала на улазу у квантизер, у

дискретан скуп пројектованих дозвољених амплитуда на излазу, а у складу са одређеним алгоритмом [2]. Посебно квантовање појединачних одмерака сигнала одликује скаларне квантизере, док је истовремена квантизација групе одмерака који формирају векторе одлика векторских квантизера [1], [3], [8]. У дисертацији се разматрају скаларни квантизери као и погодности њихове примене у алгоритмима за кодовање сигнала и параметара неуронске мреже.

У основи дигиталног кодираног сигнала је низ бинарних цифара или битова, а који уз примену различитих алгоритама за дигитално кодовање или скраћено кодовање, представљају изворне приказе сигнала са мањим бројем битова. Тако кодовани прикази се складиште и преносе до одредишта тј. пријема, на којима се помоћу алгоритама декодовања врши реконструкција оригиналног садржаја. Како се кодовањем врши редукција броја битова коришћених за приказ оригиналног садржаја, тежња је да се обезбеди што већа веродостојност реконструисаног садржаја. У том смислу кодовање представља компромис условљености за приказом компримованих верзија садржаја са редукованим бројем битова и степена квалитета репродукованог сигнала који изискује његова даља примена. Додатно се уштеда у битовима може усмерити ка алгоритмима заштите преноса и чувања садржаја, што није у фокусу ове дисертације.

Постоји широки спектар алгоритама кодовања, а сама основа тих алгоритама је у великој мери повезана са апликативним решењима у којима се користе. Генерално се алгоритми за кодовање могу класификовати на алгоритме са и алгоритме без губитака [1], [3], [4], [5]. Код алгоритама кодовања без губитака се врши идеална реконструкција информација на улазу ових кодера, што није случај код алгоритама кодовања са губицима, чија примена претходи примени алгоритама кодовања без губитака. По другој класификацији, алгоритми за кодовање се могу сврстати у три групе алгоритама за: кодовање таласног облика, параметарско кодовање и хибридно кодовање, које представља комбинацију претходна два алгоритма кодовања. Кодери таласног облика, као што и сам назив сугерише, имају примарни циљ да очувају таласни облик сигнала који се квантује/кодује, док га параметарски кодери моделују и врше реконструкцију

на основу екстрахованих параметара самог параметарског модела [2]. Поменути модели кодовања се ослањају на статистичка знања о улазним сигналима и њиховој природи. Алгоритми предложени и разматрани у дисертацији засновани су на кодовању таласног облика сигнала са губицима у поступку кодовања.

У процесу квантизације се амплитуде улазних сигнала пресликавају у амплитуде репрезентационих нивоа квантизера пројектованим за одређене битске брзине [1], [2], [9], [10]. Сама одлука о избору репрезентационих нивоа зависи од функције густине вероватноће као и од граница одлучивања квантизера, које могу бити униформне или неуниформне, што даље условљава поделу на униформне и неуниформне квантизере. У дисертацији су посебна поглавња посвећена моделима униформних и неуниформних квантизера и њиховој примени у обради сигнала и компресији неуронских мрежа.

Амплитуда максималног оптерећења одређује опсег у коме су распоређени квантизациони нивои и може се истаћи као значајан параметар за пројектовање квантизера [2]. Број репрезентационих нивоа је директно условљен битском брзином квантизера, тј. његовом резолуцијом, па је и одступање вредности амплитуда квантованих одмерака од стварних оригиналних вредности амплитуда одмерака очекивано мање уколико је резолуција већа. Грешка која се уноси квантизацијом је заправо дисторзија сигнала [1], [2], [9], [11] - [14] која се може оценити објективним или субјективним мерама. Најчешће коришћена објективна мера дисторзије је однос сигнал-квантизациони шум SQNR (*Signal-to-Quantization-Noise-Ratio*) која представља меру односа снаге улазног сигнала и шума квантизације, и иста је коришћена код компаративних анализа погодности и предности у дисертацији предложених модела квантизера. Код неуронских мрежа, као објективна мера се издваја тачност саме неуронске мреже [15], [16]. Како би се постигле што више вредности објективних мера, истиче се потреба за развојем нових алгоритама за кодовање сигнала, али и примена добро параметризованих квантизера у савременим апликативним решењима, где је потребно остварити што већи степен компресије сигнала уз незнатну деградацију тачности модела [15], [16].

Циљ ове дисертације је упознавање са методама и концептима који се користе у квантизацији неуронских мрежа и дискусија о постигнућима у овој истраживачкој области која су базирана на употреби квантизације, примарно нискобитних скаларних квантизера. Поред навођења литературе и најважнијих радова у овој области на којој је изграђена ова дисертација, посебно ће бити наглашени најрелевантнији радови коју су послужили и као мотивација за моделирање скаларних квантизера које је изведено у овој дисертацији. Препознајући истраживачку актуелност у области неуронских мрежа базирану на употреби различитих модела квантизера, као и даље тенденције развоја и прилагођења квантизера захтевима архтектуре неуронских мрежа и апликација, у овој дисертацији су дати предлози и анализе одређених модела униформних и неуниформних квантизера којима је очувана или сасвим незнатно деградирана тачност иницијалних модела неуронских мрежа. Како би истакли атрактивност предложених решења, предочени су и тренутно актуелни концепти који користе квантизацију у компресији неуронских мрежа, као и дискусија о изазовима и могућностима који се даље могу подстицати добром параметризацијом квантизера.

Прецизније, ова докторска дисертација је организована тако што су формирана посебна поглавља посвећена моделирању квантизера у области обраде сигнала и за област неуронских мрежа. Најпре су у другом поглављу дате основе скаларне квантизације, закона компресије и основне карактеристике Лапласове функције густине вероватноће. Само друго поглавље није обимно, обзиром да је квантизација стара фундаментална техника, која се користи у различитим традиционалним телекомуникационим системима, али проналази примену и у савременим системима, примарно базираним на квантованим неуронским мрежама QNN (*Quantized Neural Network*).

Предлог нових модела квантизера у области обраде сигнала и њихова параметризација дати су у трећем поглављу. Мотивисани чињеницом да униформа квантизација није најпогоднија за сигнале који имају неуниформну функцију густине раподеле, као што је Лапласова функција, у трећем поглављу се најпре анализира идеја поделе амплитудског опсега на два неједнака дисјунктна региона уз коришћење

униформних квантизера једнаких битских брзина у оба региона [17]. Конкретно, претпоставља се да CGR – централни грануларни регион, који се налази у централном делу покрива сам врх Лапласове функције густине вероватноће и део око њеног врха, док PGR – периферијски грануларни регион, покрива репове функције густине вероватноће. Оптимизација ширина PGR и CGR се базира на оптимизацији дисторзије по параметру дефинисаним односом граничног прага, који одваја CGR од PGR, и прага одсецања тј. границе грануларног региона. Показује се да резултујућа итеративна формула омогућава једноставну параметризацију део-по-део униформног квантизера PWUQ (*Piecewise Uniform Quantizer*). За средње и високе битске брзине демонстрирана је погодност описаног PWUQ у односу на униформни квантизер (UQ) уз посебну пажњу посвећену случају када 99.99 % амплитуда сигнала који се квантује припада пројектованом амплитудском опсегу квантизера тзв. региону подршке (*support region*) или грануларном региону. Резултујућа формула за PWUQ дизајн и процену перформанси је веома корисна и за неуронске мреже где су тежине типично моделоване Лапласовом функцијом густине раподеле и где се униформни квантизери често користе из разлога једноставности и како би се задовољили меморијски захтеви.

Затим је, такође у трећем поглављу, дата параметризација симетричног квантиле квантизера SQQ (*Symmetric Quantile Quantizer*) [18] предлагањем метода за тзв. *offline* калкулацију потпуно специфицираног сета параметара SQQ за улазе који се могу описати Лапласовом функцијом густине раподеле. На овај начин одређен и сачуван сет параметара се може користити у детерминистичком процесу. И у случају овако описаног модела квантизера сагледане су предности, али и ограничења квантизације, као и неминовна грешка квантизације. У наставку истог поглавља је дат квазилогаритамски квантизер са посебним осврtом на одређивање амплитуде максималног оптерећења [19], [20]. Како је квазилогаритамски квантизер робустан у широком динамичком опсегу варијансе сигнала, препознали смо посебни интерес да у зависности од коефицијента скалирања, који даје однос варијансе на улазу у квантизер и варијансе за коју је квантизер пројектован, одредимо максимални SQNR за различите битске брзине као и за различите вредности фактора компресије μ.

У четвртом поглављу је дат опис тестне архитектуре неуронске мреже и предочена је примена квантизације у фази после тренинга. Описују се MNIST (*Modified National Institute of Standard and Technology database*) и FASHION-MNIST скуп података који се у експерименталној фази истраживања доводе на улазе предложених нових или добро моделованих познатих модела униформних и неуниформних нискобитних квантизера, детаљно описаних у четвртом и петом поглављу. Од посебног интереса је сагледавање утицаја амплитуде максималног оптерећења, као и параметризације нискобитних квантизера, како на тачност неуронске мреже, тако и на тачност сваког од појединачних слојева модела неуронске мреже. За све описане моделе нискобитних квантизера [21] - [24], анализирају се и SQNR и тачност описане тестне архитектуре неуронске мреже. Указаје се на оправданост издвојене анализе за двобитне и за тробитне квантизере, обзиром на разлике у закључцима до којих се непосредно дошло, а посебно са освртом на прецизност одређивања амплитуде максималног оптерећења и потребу њеног прилагођавања стварној динамици амплитуда улазног сигнала. Главни закључци, постигнућа и допринос су сумирани и предочени у последњем, шестом поглављу ове дисертације.

2. ОСНОВЕ КВАНТИЗАЦИЈЕ

2.1 Кратак историјат квантизације

Квантизација је стара фундаментална, али и даље актуелна телекомуникациона област. Неки од ранијих прегледа историје квантизације до 1998. године дати су и раду *Gray&Neuhoff* [25]. У најширем смислу квантизација је техника мапирања улазних вредности неког великог, често скупа континуалних вредности од интереса, у мањи и коначни скуп. Још су у деветнаестом веку коришћени методи заокруживања, одсецања или дискретизације. У двадесетом веку, квантизација је била важна у дигиталној обради сигнала: говорних и аудио сигнала, слике и мултимедијалних садржаја. Сам процес представљања сигнала у дигиталном облику обично укључује заокруживање, као и имплементацију нумеричких алгоритама, где се прорачуни са реалним вредностима изводе аритметиком коначне прецизности. Средином двадесетог века, *Shannon* је дао математичку теорију комуникација [26], чиме је и формално је представљен ефекат квантизације и њена употреба у теорији кодовања. *Shannon* је предочио да је коришћење увек истог броја битова за кодовање вредности од интереса које имају различите вероватноће појављивања може бити неефикасно, те да се може применити оптималнији приступ, тј. коцепт кодовања са променљивим бројем битова или променљиве битске брзине. У даљем раду *Shannon* је увео појам векторске квантизације која је нашла практичну примену у реалним комуникационим апликацијама. У области обраде сигнала важан моменат представља увођење *PCM* (*Pulse Code Modulation*) [27], пулсирајуће методе предложене за кодовање одмерених аналогних сигнала, као и квантизација високе резолуције [10].

Квантизација се појављује у нешто другачијој форми у алгоритмима који користе нумеричку апроксимацију за проблеме који укључују континуалне математичке величине, тј. области која такође има дугу историју, али је добила ново интересовање са појавом дигиталног рачунара. У нумеричкој анализи важан појам је добра поставка проблема. Може се рећи да је проблем добро постављен или дефинисан уколико постоји јединствено решење које непрекидно зависи од улазних података у некој

топологији или архитектури. Овакви се проблеми понекад називају добро условљеним проблемима. Испоставило се да, чак и када се ради са датим добро условљеним проблемом, одређени алгоритми који тачно и прецизно решавају тај проблем у неком идеализованом смислу, могу да раде лоше у присуству дисторзије која уноси грешаке заокруживања и одсецања, уопштено квантизационе грешке. Саме грешке заокруживања, које имају везе са представљањем реалних бројева коначним бројем битова и коришћењем презентација са покретном тачком, као и грешке одсецања, настају и будући да се изводи коначан број итерација самог итеративног алгоритма.

2.2 Основни појмови у скаларној квантизацији

Према општој подели квантизери се могу класификовати на скаларне и векторске квантизере. Код скаларних квантизера се квантује одмерак по одмерак, док се код векторских квантизера врши груписање одмерака који се затим, као група истовремено квантују. У овој дисертацији се пажња посвећује скларaним квантизерима. У даљој класификацији, скаларни квантизери су подељени на униформне, неуниформне и део- по-део униформне квантизере. Наведеним врстама скаларних квантизера биће посвећена посебна пажња у наредним поглављима, уз сагледавање предности примене сваког од наведених модела квантизера у областима од интереса.

У традиционалним телекомуникационим решењима, улазни сигнал је добро моделован неком од познатих функција густине вероватноће која најбоље описује сам сигнал на улазу квантизера који је пројектован и прилагођен управо тој функцији густине вероватноће.

Лапласов извор без меморије са средњом вредношћу μ и варијансом σ^2 дефинисан је Лапласовом функцијом густине вероватноће [1], [3], [4]:

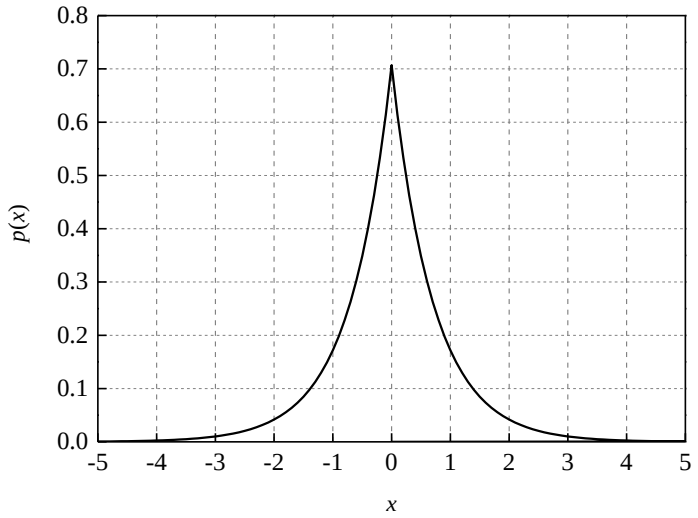
$$p(x)=\frac{1}{\sqrt{2}\sigma}\exp\left\{-\frac{\sqrt{2}|x-\mu|}{\sigma}\right\}, \tag{2.2.1}$$

док је Гаусов извор средње вредности μ и варијансе σ^2 дефинисан је као:

$$p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{2.2.2}$$

Иако у изразима (2.2.1) и (2.2.2) фигурише и средња вредност μ, у анализама модела квантизера у овој дисертацији користиће се нулта средња вредност, док ће сама варијанса бити увек експлицитно наведена. Наиме, предложићемо моделе униформних и неуниформних квантизера који подразумевају симетрију дизајна.

Према [28], Лапласова функција густине вероватноће (видети слику 2.2.1) је најприкладнији облик за моделирање статистичких својстава аудио и говорних сигнала. Као што је познато, Лапласова и Гаусова функција густине вероватноће припадају породици лог-конкавних функција. Концепт лог-конкавности био је широко проучаван у литератури за многе симетричне функције густине вероватноће, попут Лапласове, као у [29]. Лог-конкавност може олакшати тачно предвиђање дизајна квантизера са становишта његове симетрије. Према [29], како Лапласова функција густине вероватноће задовољава услов лог-конкавности, оптимални квантизер је и сам симетричан. Конвексност MSE дисторзије недавно је добила велику пажњу у [14].



Слика 2.2.1. Лапласова функција густине вероватноће нулте средње вредности и јединичне варијансе.

Недостатак конкавности виђен је и доказан за непарне бројеве репрезентационих нивоа квантизера, док је за било коју симетричну функцију густине вероватноће, попут Лапласове, MSE (*Mean-Squared Error*) дисторзије, тј. средње-квадратна грешка дисторзије, је конвексна за паран број репрезентационих нивоа [14].

Основни параметри којима се квантизери добро описују су:

- Битска брзина;
- Границе (нивои) одлучивања;
- Репрезенти (репрезентациони нивои);
- Амплитуда максималног оптерећења.

За моделе квантизера, описане у овој дисертацији, на почетку сваког одељка посвећеног том квантизеру биће дата дефиниција самог квантизера и квантизационих ћелија. Затим ће следити општа дефиниција основних параметара анализираног модела квантизера и вршиће се оптимизација истих, док су остварена перформансна побољшања приказана у одељцима са експерименталним и нумеричким резултатима.

Познато је да се при квантизацији реална оса, на којој су прадстављени улази квантизера, дели на квантизационе ћелије или квантизационе интервале, у ознаци Δ_i, који могу бити униформно или неуниформно распоређени, тј. могу бити једнаких или неједнаких ширина, те се и сами квантизери по овом основу могу поделити на униформне и неуниформне. Границе квантизационих ћелија су нивои одлучивања и њима се одређује припадност одмерка улазног сигнала одређеној квантизационој ћелији, те се свим одмерцима из те квантизационе ћелије додељује репрезент дефинисам мапирањем или пресликавањем квантизера. Како се вредност репрезента дате квантизационе ћелије разликује од реалне вредности улазног сигнала, квантизационим процесом се уноси неизбежна грешка квантизације. Усредњена сума свих квантизационих грешака квантизера за дати сигнал на улазу представља дисторзију. Будући да многи реални сигнали могу показивати различите степене стационарности, често је повољно омогућити флексибилне кодере подржане

флексибилним брзинама преноса, који се у реалном времену могу прилагодити атрибутима сигнала који се стално мењају у времену.

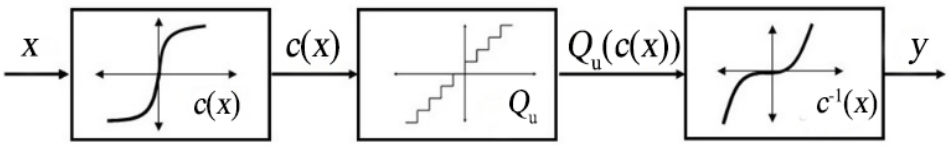
Дозвољене вредности амплитуда улазног сигнала се налазе унутар грануларног региона квантизера који је једнозначно оређен амплитудом максималног оптеређења квантизера x_{max}. У овој дисертацији разматраћемо различите моделе симетричних квантизера, чији је грануларни регион дат као $[-x_{max}, x_{max}]$. Овако дефинисан опсег максималних амплитуда одређује регион подршке или грануларну област квантизера (*granular region*). Сви одмерци улазног сигнала који не припадају грануларној области су додељени тзв. области прекорачења (*overload region*). У складу са дефиницијом дисторзије, укупна дисторзија се може наћи као сума грануларне дисторзије и дисторзије прекорачења. По одређивању свеукупне дисторзије, одређује се SQNR, као објективна мера која омогућава упоредивост описаног модела квантизера, за дати улаз, са другим моделима квантизера, за исти улаз.

Недавно су дате анализе *Sangsin Na et al.* [12], [30] за најједноставнији модел квантизера, то јест за униформни квантизер. Рад [12] бави се монотоношћу по два за симетричне униформне квантизере у случају четири најчешће функције густине вероватноће, укључујући и Лапласову функцију густине вероватноће, коју посебно анализирамо у овој дисертацији. Још једна занимљива чињеница, наведена и доказана за Лапласову функцију густине вероватноће у овим радовима јесте да се за довољно велики број нивоа квантизације, оптимална величина корака за униформни квантизер смањује, како се битска брзина повећава. Иницијално изведена из "Теореме о делимичним дисторзијама" или *Partial Distortion Theorem* (PDT), анализа дата у [30] даје неколико увида у однос између MSE дисторзије, броја и положаја нивоа квантизације у унутрашњем и спољашњем региону, или грануларном и региону прекорачења.

Генерално, да би се оптимизовао модел квантизера, потребно је познавање статистике и функције густине вероватноће улазног сигнала, уз прилагођење статистици самог сигнала на најбољи могући начин. Симетрична Лапласова функција густине вероватноће сигнала са истакнутим врхом и дебелим реповима се користи за

ефикасно моделовање сигнала у бројним апликацијама [28], [31], [32]. Штавише, то је вероватно најпогоднија форма функције густине вероватноће за говорне и аудио сигнале јер веома добро описује бројне карактеристичне атрибуте ових сигнала [28], [30], [33]. Додатно, трансформисани сигнали и остали квантитети изведени из оригиналних сигнала, често прате Лапласову функцију густине вероватноће [28]. Као што се често среће у многим апликацијама, и у овој дисертацији се фаворизује моделовање сигнала помоћу Лапласове функције густине вероватноће.

У овом одељку се даље за класу неуниформних квантизера дефинише и компресорска функција $c(x)$. Наиме, једна класа неуниформних квантизера (NUQ), тзв. компандинг квантизера (*companding quantizers*) се пројектује применом A или μ закона компресије. Сам неуниформни компандинг квантизер се може представити редном везом компресора, униформног квантизера и експандора (видети сл. 2.2.2.). Улазни сигнал је компримован применом компресорске функције $c(x)$, а затим је доведен на улаз униформног квантизера. Квантовани сигнал се доводи на улаз експандора, који применом инверзне компресорске функције $c^{-1}(x)$, генерише излазни сигнал y.



Слика 2.2.2. Модел компандора.

Код компандора дефинисаних A законом компресије, компресорска функција има облик [3] - [6]:

$$c_A(x)=\begin{cases}\dfrac{A|x|}{1+\ln A}\operatorname{sgn}(x), & 0\le\dfrac{|x|}{x_{\max}}\le\dfrac{1}{A}\\[2ex] \dfrac{x_{\max}}{1+\ln A}\left(1+\dfrac{A|x|}{x_{\max}}\right)\operatorname{sgn}(x), & \dfrac{1}{A}\le\dfrac{|x|}{x_{\max}}\le 1\end{cases}, \qquad (2.2.3)$$

где A представља параметар компресије, дефинисан приликом пројектовања квантизера, док sgn(x) је *signum* функција и представља знак тренутне вредности улазног сигнала.

Код компандора дефинисаних μ законом компресије, компресорска функција има облик [3] - [6]:

$$c_{\mu}(x)=\frac{x_{\max}}{\ln(1+\mu)}\ln\left(1+\mu\frac{|x|}{x_{\max}}\right)\operatorname{sgn}(x),\quad -x_{\max}\le x\le x_{\max}. \tag{2.2.4}$$

Многи напори у класичној компресији уложени су у правцу минимизације неизбежне грешке квантизације, а за дату брзину преноса, обзиром да је главни циљ компресије заправо редуковање брзине преноса. Међутим, пошто генерално важи – што је битска брзина нижа, то су мањи меморијски захтеви, али је већа грешка квантовања [2], због ових конфликтних захтева, квантизација је веома интригантна област истраживања. Конкретно, избор самог модела квантизације и спецификација његових кључних параметара: {грануларног региона, величине квантизационих ћелија, битске брзине, нивоа одлуке и репрезентације} [12] - [25], [34], утиче на укупну грешке квантизације, односно тоталну дисторзију. Пратећи главни циљ кодовања и компресије сигнала, а то је смањење брзине преноса или њено усклађивање са апликативним захтевима, можемо очекивати да за неуниформне изворе, као што је Лапласов, који је претпостављен у анализама у овој дисертацији, неуниформна квантизација омогућава боље коришћење доступне битске брзине. Додатно ограничење, посебно за услове ниских битских брзина, представља прихватљива сложеност дизајна и имплементације неуниформних квантизера.

2.3 Примена квантизације у неуронским мрежама

Људи су јединствени по својој способности да одмере перцепције и доносе одлуке на основу њих. Људско расуђивање је инхерентно субјективно и стога се разликује од појединца до појединца. Осим тога, емоционални, когнитивни и друштвени процеси могу надјачати основне инстинкте и условити избор понашања и одлука. У том погледу, људско закључивање је последица интеракције свих ових фактора. У пракси није реално идентификовати и адекватно квантификовати све факторе који играју улогу у одређеном феномену, чак и ако феномен делује једноставно. Како количина података која нам је потребна за генерализацију расте, тако расте и број функција или димензија. Под претпоставком да је објашњавајући део многих проблема окарактерисан као високодимензионалан, како би се подаци фиксирали у линеарном простору или на неком простору знатно мање димензије, пожељно је укључити неку од познатих функција густина вероватноће. Често се у литератури претпоставља да су функције густине вероватноће сигнала познате, тј. да сигнали прате неку од основних расподела [28], [29], [31]. Међутим, у многим ситуацијама обично нема претходних информација о расподели сигнала који се обрађује, тако да се усваја широко прихваћено мишљење да Лапласова функције густине вероватноће добро описује многе природне, економске и друштвене појаве [31]. Уопштено говорећи, као последица делимичног или потпуног недостатка информација о функцији густине вероватноће, прихватање неке од расподела, попут Лапласове расподеле, нуди могућност потпунијег сагледавања атрибута улазних сигнала и њиховог међусобног односа. То оправдава наш избор о усвојању Лапласове функције густине вероватноће у анализама датим у овој дисертацији.

Многа савремена решења су инхерентно децентрализована или хетерогена и подаци се преносе на мрежном нивоу без агрегације због различитих ограничења комуникационих система [35]. Тенденција је да се преносе компримовани или квантовани подаци пре него сирови и необрађени [15], [16]. Како је квантизација виђена као пожељан механизам за редукцију броја битова, квантизери имају значајну улогу у оптимизацији или побољшању постојећих мрежних решења. Бројне

индикације потврђују да се погодним моделовањем квантизера може значајно редуковати количина пренетих информација или података од интереса. Како се исти квантизери користе за улазне сигнале који имају различите и променљиве статистике, постизање робустности моделованих квантизера може бити јако битно.

Приликом анализе ефикасности квантизера, пожељно је да се модели квантизера и функције густине вероватноће приближе стварним подацима. Иако је могуће постићи високу тачност на стварном скупу обуке, прави изазов је развити модел који добро генерализује податке који нису познати унапред. Наведени критеријум за модел квантизера логично води до решења које дозвољава одређеним количинама података да диктирају алгоритме и решења која избегавају недовољно прилагођавање и претерано прилагођавање модела. Уопштено говорећи, присуство прекомерног прилагођавања уобичајен је проблем многих модела неуронских мрежа [36]. Осим тога, није могуће делимично спречити настанак могућих грешака екстраполације, које се обично појављују у непокривеним регионима, без података. Стога се намеће се основно питање како изградити робусне системе који разумно и безбедно раде у стварном свету. Постизање робусности свакако је циљ не само за прецизну неуронску мрежу, већ и за квантовану неуронску мрежу [15], [37]. Инспирисани заједно оптимизацијом броја бита и робусношћу, истраживање и тумачење различитих принципа квантизације који стоје иза квантоване неуронске мреже, остају интригантни правци за будућа истраживања.

Вођени потребом за компресијом параметара неуронске мреже, што је посебно корисно за меморијски ефикасну примену неуронске мреже на уређајима са ограниченим ресурсима, бројни радови су већ потврдили обиље могућности за компресију параметара неуронске мреже помоћу квантизације [37] - [50]. Процес квантовања уводи неизбежну грешку квантовања [8], [11], [14], а њено акумулирање може проузроковати погрешан излаз квантоване неуронске мреже, чиме се деградира тачност у поређењу са неуронским мрежама. Пошто је грешка квантизације индикатор квалитета сваког квантизера, од посебног је интереса испитати њену везу са тачношћу квантоване неуронске мреже.

Генерално, у нискобитној квантизацији, веома мали број бита по одмерку се користи за представљање података који се квантују (мање или једнако 3 бит/одмерак) [2]. Ослањајући се на мноштво модела квантизације из области обраде сигнала, квантизација се показала као ефикасна техника која може извршити компресију сигнала према неким од основних критеријума. Нови приступи компресији сигнала, који укључују коришћење квантизера у моделима неуронских мрежа, отворили су бројне могућности коришћења квантизера у системима и апликацијама из реалног живота. Сфере примене неуронских мрежа су многоструке, а обухватају између осталог и препознавање и класификацију слика, као и обраду и препознавање природног говора [15], [16]. Побољшање тачности модела неуронских мрежа је често праћено повећањем самих модела чинећи их превише параметеризованим, али и неадекватним за имплементацију у системима са ограниченим ресурсима. Постизање ефикасних неуронских мрежа у реалном времену са жељеном тачношћу захтева преиспитивање дизајна, обуке и примене модела неуронских мрежа [37]. Зато је значајна пажња усредсређена на изналажење решења која би моделе неуронских мрежа учиниле ефикаснијим у смислу кашњења, меморијског простора, утрошка енергије, хардверске и рачунарске комплексности итд.

Пројектовање ефикасних неуронских мрежа се може сагледати кроз различите аспекате који укључују [16]:

а) Оптимизацију архитектура неуронске мреже: Док су класичне методе оптимизације базиране на мануелном претаживању нових архитектура, новије методе дизајнирања неуронске мреже за аутоматизовано машинско учење и претраживање неуронске архитектуре (NAS) имају за циљ да на аутоматизовани начин пронађу адекватну архитектуру неуронске мреже, под датим ограничењима величине, дубине или ширине модела [51].

б) Редизајн архитектуре неуронских мрежа и хардвера: Важност овог здруженог редизајна неуронских мрежа се огледа у оптимизацији неуронских мрежа имајући у виду да харвдерске компоненте значајно утичу на кашњење, утрошак енергије, као и на рачунарску комплексност [52].

в) Орезивање (*Pruning*): Овај приступ подразумева уклањање неурона мале активности (*sensitivity*) уз праћење утицаја на функцију модела неуронских мрежа и евентуалног губитка укупне тачности. Орезивање неурона мале активности се може вршити селективним али и агресивним уклањањем неурона мале активности, што незнатно афектира генерализационе перформансе модела неуронских мрежа и потпада под неструктуирано орезивање [53]. Уколико се уклања група параметара (нпр. читави конволуциони филтри) значајно се мењају улазни/излазни слојеви и тежинске матрице и врши се структуирано орезивање које може резултовати значајном деградацијом тачности [54].

г) Дестилацију знања: Дестилација модела укључује обуку великог модела, а затим употребу тог модела као „наставника“ за обуку компактнијег модела „ученика“ [55]. Како би постигла виши степен компресије, дестилација знања се комбинује са другим приступима нпр. орезивањем и квантизацијом.

д) Квантизацију: Коришћење квантизације је показало велики и доследан успех доминантно у фазама закључивања, пост-тренинга и поновног тренирања неуронске мреже, док је примена квантизације у фази обуке модела неуронске мреже захтевнија и осетљивија по питању тачности модела неуронских мрежа [40]. Конкретно, увођење целобројних, полупрецизних, мешовито-прецизних или нижебитних представа пуне прецизности [44] главни су концепти који се користе код квантованих неуронских мрежа, а који су предмет бројних дискусија и истраживања о изазовима и могућностима примене квантизације код квантованих неуронских мрежа.

ђ) Квантизацију неуронских мрежа и неуронауку: Иако постоји слаба веза квантизације неуронских мрежа са неуронауком, за неке ауторе је то управо мотивишућа веза, обзиром да се сматра да људски мозак складишти информације у дискретном/квантованом облику, а не у континуалном аналогном облику [57]. Једно од образложења ове поставке јесте управо осетљивост на буку, обзиром да би бука неизбежно покварила информације ако би се исте чувале у континуалном тј. аналогном облику.

У области квантизације су написани бројни радови, а многи од скорашњих радова у области неуронских мрежа који се базирају на примени постојећих техника и метода квантизације, укључују и неке „*нове моделе*“ добро прилагођених квантизера. Саме неуронске мреже доносе бројне изазове и могућности, те се тиме проширују и могућности примене различитих модела квантизера.

Неуронске мреже најчешће одликује висока архитектурна, меморијска и рачунска комплекност. Већина актуелних модела су превише параметризовани те је циљ не само редукција архитектуре мреже него и укупног броја параметара мреже. Добра околност неуронских мрежа јесте управо отпорност чак и на агресивну квантизацију и екстремну дискретизацију, што ствара додатни степен слободе у побољшању њихових перформанси путем квантизације.

У области неуронских мрежа не постоји ниједан добро постављен и условно засигурно решен проблем. Наиме, имајући у виду високу комплексност и зависност саме неуронске мреже од мноштва улазних података којима се описује у општем случају проблем или задатак од интереса, постоји много веома различитих модела који тачно или приближно оптимизују дизајн неуронске мреже. Такође је потребно нагласити да постоји и могућност постојања велике грешке или разлике квантованог модела од оригиналног (неквантованог) модела, што ипак има за исход врло добре перформансе генерализације. Наиме, овај додатни степен слободе није био присутан у многим класичним истраживањима, која су се углавном фокусирала на проналажење метода компресије чијом применом сам сигнал који се компримује не би био превише промењен, а коришћене нумеричке методе вршиле су строгу контролу разлике између стварног, тј. оригиналног и квантованог сигнала. Управо ова строгост захтева који постоје код квантизације традиционалних телекомуникационих мрежа, послужили су као добра основа и запажање које је мотивисало истраживање нових форми квантизације у области квантованих неуронских мрежа. Коначно, слојевита структура модела неуронских мрежа нуди додатну димензију за бројна истраживања и моделовања. Различити слојеви у неуронској мрежи имају различит утицај на свеукупну тачност, што мотивише приступ квантизације мешовите прецизности [44].

Редукција свеукупне мрежне комплексности која се постиже квантизацијом, не условљава нужно и последично редукцију свеукупне тачности [16], [41].

Квантизација се посебно показала као ефикасна код имплементације неуронских мрежа у уређајима ограничених ресурса, пошто се целокупна неуронска мрежа може имплементирати у *on-chip* меморију крајњег (*edge*) уређаја, те се тако може редуковати *overhead* због приступа off-chip меморији [16]. Наиме, стандардна имплементација неуронске мреже подразумева 32-битну пуну прецизност (*FP32*) презентације параметара неуронске мреже, захтевајући комплексан и скуп хардвер. Подробним избором квантизера за модел неуронске мреже, тј. за квантовање тежина и активација и представом са ниском прецизношћу, може се значајно редуковати жељена битска брзина за дигиталну представу параметара неуронске мреже, уз смањење меморијске комплексности и одржавање тачности модела неуронске мреже [38], [40], [50]. Из тог разлога бројни нови модели квантизера и методологија квантизације су предложени уз задовољење главног циља оптимизације квантованих неуронских мрежа, а то је - незнатна деградација тачности модела или у најбољем случају, одржавање тачности модела неуронске мреже.

2.4 Конкретна примена нискобитних скларних квантизера у неуронским мрежама

Од четвртог поглавља тражићемо ефикасна решења која укључују примену квантизације у фази после тренинга неуронске мреже, а из перспективе избора амплитуде максималног оптерећења при примени нискобитне квантизације. Даћемо експерименталне и теоријске резултате за неколико значајних случаја дизајна униформних и неуниформних квантизера, где претпостављамо да Лапласов извор моделира расподелу тежина у нашем потпуно повезаном моделу неуронске мреже. Сама Лапласова функција густине вероватноће доказано најбоље моделира расподелу тежина неуронске мреже [16]. Такође ћемо анализирати да ли је могуће применити најједноставнију униформну квантизацију за представљање тежина обученог модела неуронске мреже са битском брзином од $R = 2$ бит/одмерак, уз очување тачности модела у великој мери. За циљ постављамо утврђивање избора кључног параметра двобитног униформног квантизера, тј. амплитуде максималног оптерећења и показујемо како се избор овог кључног параметра одражава на SQNR и тачност модела неуронске мреже. Штавише, проширићемо анализу на случај двобитне униформне квантизације по слојевима како бисмо испитали да ли је могуће постићи додатно побољшање тачности предложеног модела неуронске мреже за MNIST скуп података. Сматрамо да је детаљна анализа пост-тренинг квантизације описана и спроведена у четвртом поглављу веома корисна за сва даља истраживања ове веома актуелне теме, посебно због чињенице да се проблем пост-тренинг квантизације разматра из посебно важне перспективе избора амплитуде максималног оптерећења. Квантизација неуронске мреже након обуке, уколико се сачува тачност, може бити погодна јер не захтева поновно обучавање и фино подешавање неуронске мреже, док се величина меморије потребна за чување тежина модела квантоване неуронске мреже може значајно смањити у поређењу са основним моделом неуронске мреже, који за њихово чување користи 32-битни FP32 формат.

У четвртом поглављу смо показали да тачност два модела неуронске мреже MLP (*MultiLayer Perceptron*) и CNN (*Convolutional Neural Network*), које смо претходно обучили за MNIST скуп података, може бити очувана за различите изборе кључног параметра, тј. амплитуде максималног оптерећења, када се тробитни униформни квантизер користи за квантизацију тежина неуронске мреже у пост-тренинг фази. Деградација тачности је процењена у односу на иницијалну тачност трениране неуронске мреже са тежинама представљеним у FP32 формату, који је подразумевани формат за GPU (*Graphics Processing Unit*) и CPU (*Central Processing Unit*) платформе. Такође смо показали да у тробитној униформној квантизацији тежина промењеној у пост-тренинг фази, за оба модела неуронске мреже (MLP и CNN) и оба сета (MNIST и Fashion-MNIST), није од највеће важности, као што је то у класичној квантизацији, одређивање оптималне вредности амплитуде максималног оптерећења униформног квантизера, да би се постигла нека унапред дефинисана тачност квантоване неуронске мреже. За разлику од случаја са двобитном униформном квантизацијом, где смо интуитивно предвидели и показали да избор ширине грануларне области има велики утицај на тачност модела квантоване неуронске мреже, у овом поглављу, а за исти класификациони задатак и исто специфицирани MLP, нисмо дали такво предвиђање у случају да униформни квантизер има само један додатни бит, па смо испитали утицај избора ширине грануларног региона за тробитни униформни квантизер на тачност квантоване неуронске мреже и утврдили слабу зависност тачности неуронске мреже од амплитуде максималног оптерећења, нарочито за MLP модел. Посебно смо истакли да неодговарајући избор амплитуде максималног оптерећења двобитног униформног квантизера може значајно да деградира тачност квантоване неуронске мреже, док је подешавање вредности амплитуде максималног оптерећења далеко једноставније и мање строже у случају са тробитним униформним квантизером. Показали смо да смо коришћењем тробитног униформног квантизера уз компримовање тежина MLP и CNN (тренираних на MNIST сету) више од 10 пута, успели да значајно очувамо тачност модела неуронске мреже, који је деградиран за 0.13 % и 0.10 %, респективно. Једноставност нашег предлога, заједно са високом робусношћу тачности на промену вредности амплитуде максималног оптерећења, указује на то да се предочени модел

квантизације у пост-тренинг фази може искористи на једноставан начин, што је посебно важно код ивичних уређаја ограничених капацитета. Другим речима, пошто се истраживање перформанси нискобитне квантизације може сматрати атрактивном и моћном анализом примене техника компресије, која би могла да омогући уклапање квантоване неуронске мреже у ивични уређај са прецизношћу која је очувана у великој мери, може се очекивати да ће анализа обављена у четвртом поглављу имати и велики практични значај.

У поглављу посвећеном неуниформним квантизерима предложена су се два нова модела двобитних неуниформних квантизера, који врло промишљено примењују једну од две особине најједноставнијег униформног квантизера. Такође смо испитали да ли су ови модели квантизера погоднији за квантизацију у пост-тренинг фази од униформног квантизера. Оптимизацијом дисторзије предложених неуниформних квантизера, а да бисмо постигли највиши теоријски SQNR, извели смо формуле за итеративно одређивање основних величина корака оба неуниформна квантизера, Δ и Δ^{mod}, што је од највеће важности у традиционалној квантизацији. Доказали смо да су коришћени итеративни алгоритми дали вредности Δ и Δ^{mod} које су заиста оптималне за описане моделе неуниформних квантизера, јер се поклапају са одговарајућим резултатима нумеричке оптимизације дисторзије по овим базичним амплитудним квантима. Такође смо потврдили премису да квантизер који обезбеђује највећи SQNR не мора нужно да обезбеди највећу тачност квантоване неуронске мреже. Коначно, увођењем другог, модификованог модела неуниформног квантизера, успели смо не само да побољшамо SQNR, већ и да повећамо тачност квантоване неуронске мреже у поређењу са случајем где су имплементирани униформни квантизер и први немодификовани модел неуниформног квантизера. Иако је униформни квантизер веома експлоатисан модел квантизера због своје једноставностти, али и интригантне природе и могућности да се донекле модификује, природно је очекивати да ће истраживање и развој неких модификованих модела квантизације наставити да привлаче пажњу научне заједнице.

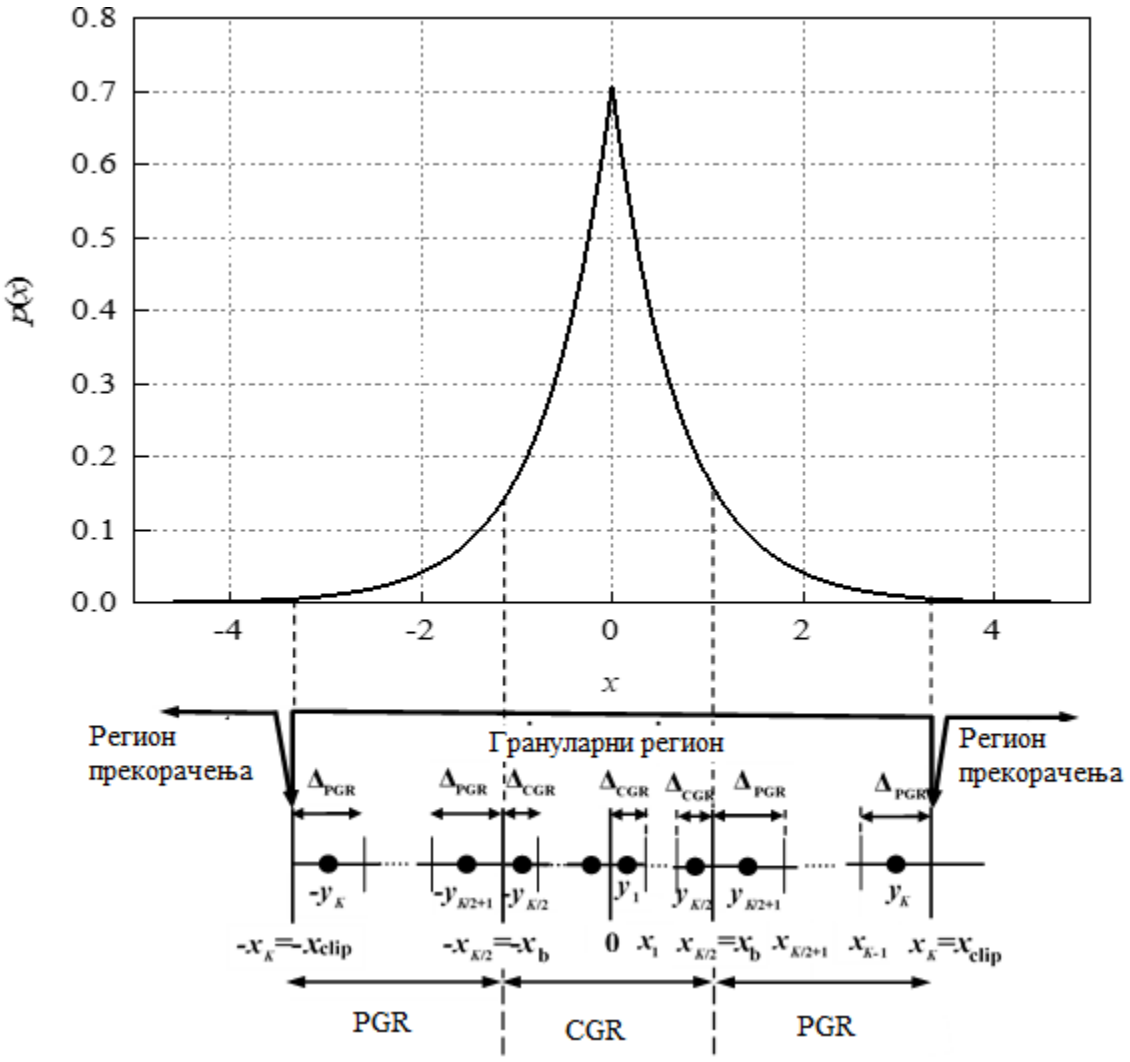
3. Пројектовање нових модела квантизера и њихова примена у обради сигнала

3.1 Итеративни алгоритам за параметризацију дво-регионалног део-по-део униформног квантизера оптимизованог за Лапласов извор

Познато је да модел неуниформног квантизера, NUQ, који је добро прилагођен амплитудској динамици улазног сигнала описаног неуниформном функцијом густине вероватноће, даје мању грешку услед квантизације у поређењу са моделом униформног квантизера, UQ, са истим бројем квантизационих нивоа или са истом битском брзином, а за исти улазни сигнал [2], [12] - [14], [17]. Међутим, због чињенице да је униформни квантизер најједноставнији модел квантизера, он је и најчешће коришћен у компресији неуронских мрежа, као што је истакнуто у [21]-[24], [39] - [42]. Штавише, висока комплексност неуниформног квантизера потенцијално може бити отежавајућа околност у постизању перформансне предности над униформним квантизерима. Сам део-по-део униформни квантизер се по комплексности налази између униформних и неуниформних квантизера. У овом одељку описујемо дизајн нашег дво-регионалног део-по-део униформног квантизера (PWUQ), састављеног од пара оптимизованих униформних квантизера са истим битским брзинама, који се у односу на униформни квантизер, боље прилагођавају статистици улазног сигнала са Лапласовом функцијом густине вероватноће, у унапред дефинисаним амплитудским регионима. Затим вршимо параметризацију PWUQ уз детаљан приказ перформансних побољшања у односу на решења са униформним квантизерима.

3.1.1 Пројектовање квантизера и развој алгоритма

Из [2] и [57] можемо усвојити дефиницију грануларног региона и раздвајање амплитуда улазног сигнала у грануларни регион и регион прекорачења, или унутрашњи и спољашњи регион, респективно. За симетричне квантизере, као и за модел који описујемо у овом одељку, ова два региона су разграничена амплитудом максималног оптерећења или прагом одсецања, означеним са $\pm x_{clip}$ (сл. 3.1.1.1.).



Слика 3.1.1.1. Подела грануларног региона на CGR и PGR: Δ_{CGR} и Δ_{PGR} су ширине ћелија у $\Re^{CGR}$ и $\Re^{PGR}$, x_i, $i = 1,\ldots,K$ су ненегативни нивои одлуке, y_i, $i = 1,\ldots,K$ су ненегативни репрезентациони нивои.

Ови прагови имају реалне вредности и дефинишу грануларни регион квантизера $[-x_{clip}, x_{clip}]$, где је дисторзија због одсецања и квантизације ограничена. Како је Лапласова функција густине вероватноће у општем случају неограничена, значајан проценат одмерака је концентрисан око средње вредности, а само мали део одмерака у грануларном региону, а близу амплитуде максималног оптерећења или изван грануларног региона.

Треба приметити да сужавање грануларног региона за фиксан број квантизационих нивоа резултира редукцијом грануларне дисторзије, док се истовремено уочава нежељено повећање дисторзије прекорачења [12]. С друге стране, за дати број квантизационих нивоа, са порастом ширине грануларног региона, регион прекорачења као и дисторзија прекорачења су редуковани уз нежељени пораст грануларне дисторзије. Главни компромис у дизајну квантизера је управо прилагођење ширине грануларног региона, који је довољно широк да се добро прилагоди амплитудској динамици сигнала, док је с друге стране, довољно узак тако да се минимизира укупна дисторзија.

Анализирајући специфичности Лапласове функције густине вероватноће, у овом одељку описујемо нови PWUQ модел [17] чији је грануларни регион довољне ширине, тако да се регион прекорачења може занемарити. Сам грануларни регион погодно је подељен на два дисјунктна региона CGR (*Central Granular Region*) и PGR (*Peripheral Granular Region*), при чему се у оба региона користи најједноставнији униформни скаларни квантизер. Основна полазна претпоставка је ужи CGR око средње вредности, која покрива врх Лапласове функције густине вероватноће, и шири PGR у остатку грануларног региона, где су репови саме Лапласове функције густине вероватноће. Ова два региона су разграничена праговима означеним са $\pm x_b$, који су симетрични у односу на средњу вредност. Како је циљ минимизирати свеукупну дисторзију, доминантно грануларну дисторзију, параметризација PWUQ модела тако да већина одмерака (99.99 %) претпостављене Лапласове функције густине вероватноће припада GR, се може сагледати као свеобухватни оптимизациони задатак. Штавише, у наставку

овог одељка ће бити прецизирано како се жељена параметризација може на погодан начин реализовати коришћењем итеративног поступка.

Најпре ћемо се позвати на основну теорију о униформним квантизерима. Наиме, униформни квантизер са N-нивоа у ознаци Q_N је дефинисан мапирањем $Q_N : \mathbb{R} \to Y$ [2], где је $\mathbb{R}$ скуп реалних бројева, $Y \equiv \{ y_{-N/2}, \ldots, y_{-1}, y_1, \ldots, y_{N/2} \} \subset \mathbb{R}$ је кодна књига величине N која садржи репрезентационе нивое y_i, тако да је задовољен услов $N = 2^r$, где је r битска брзина. Применом квантизера са N нивоа Q_N, $\mathbb{R}$ је подељен у N ограничених грануларних ћелија $\Re_i$ и две неограничене ћелије у области прекорачења. Нотација коришћена за i-ту грануларну ћелију је $\Re_i = \{x \mid x \in [-x_{\text{clip}}, x_{\text{clip}}],\ Q_N(x) = y_i\}$, тако да важп $\Re_i \cap \Re_j = \emptyset$, за $i \neq j$. Другим речима, y_i специфицира i-ту кодну реч и једини је представник за све реалне вредности x из $\Re_i$.

Симетрични PWUQ из [17], се састоји од два униформна квантизера са истим бројем квантизационих нивоа $N/2 = K$. Један квантизер се користи за квантовање амплитуда које припадају CGR $[-x_b, x_b]$ док се други користи за PGR $[-x_{\text{clip}}, -x_b) \cup (x_b, x_{\text{clip}}]$ (видети сл. 3.1.1.1.). Претпоставимо такође да су амплитуде које припадају ћелијама прекорачења, тј. интервалима. $(-\infty, -x_{\text{clip}}) \cup (x_{\text{clip}}, +\infty)$, одрезане (*clipped*).

Даље дефинишемо CGR и PGR као:

$$\begin{aligned} \Re^{\text{CGR}} &= \bigcup_{i=-K/2}^{-1} \Re_i \cup \bigcup_{i=1}^{K/2} \Re_i = \left[x_{-K/2}, x_{K/2}\right], \\ \Re^{\text{PGR}} &= \bigcup_{i=-K}^{-K/2-1} \Re_i \cup \bigcup_{i=K/2+1}^{K} \Re_i = \left[x_{-K}, x_{-K/2}\right) \cup \left(x_{K/2}, x_K\right] \end{aligned}, \tag{3.1.1.1}$$

при чему су ова два грануларна региона разграничена прагом означеним са $\pm x_b$, тако да важи $\pm x_b = \pm x_{K/2}$.

Због симетрије, Лапласова функција густине вероватноће нулте средње вредности и јединичне варијансе $\sigma^2 = 1$ за коју пројектујемо PWUQ, дата је изразом:

$$p(x) = \frac{1}{\sqrt{2}\sigma} \exp\left\{ -\frac{\sqrt{2}|x|}{\sigma} \right\} \Bigg|_{\sigma^2=1} = \frac{1}{\sqrt{2}} \exp\left\{ -\sqrt{2}|x| \right\}. \tag{3.1.1.2}$$

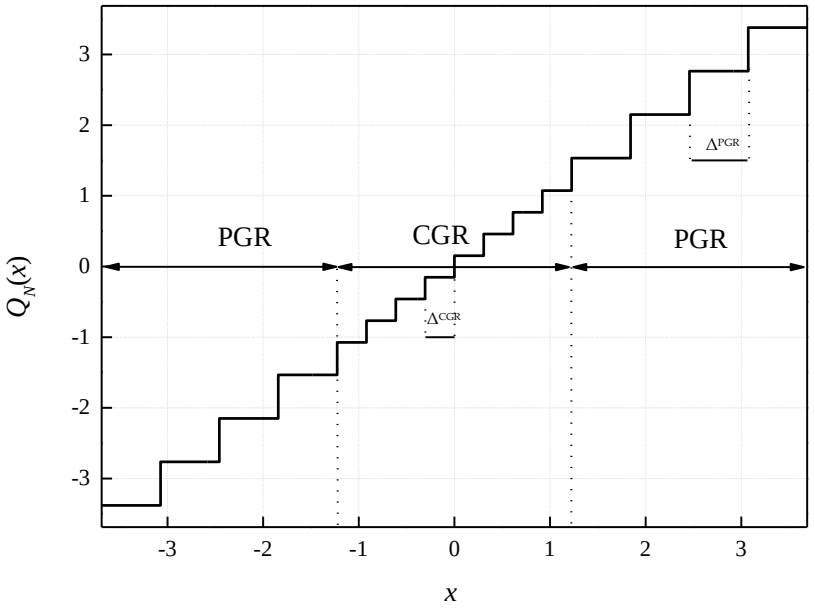
док су репрезентациони нивои и нивои одлуке симетрично позиционирани у односу на средњу вредност. Не умањујући општост, фокусирамо се на K позитивних ћелија, док за ћелије у негативном делу функције густине вероватноће тривијално следе закључци из симетрије саме функције:

$$x_{-i} = -x_i, i = 1,2,...,K \ . \tag{3.1.1.3}$$

Ненегативни нивои одлуке овог модела PWUQ су специфицирани као:

$$x_i = \begin{cases} i \cdot \dfrac{2\,x_{\mathrm{b}}}{K}, i = 0,...,\dfrac{K}{2} \\ \dfrac{(2i-K)\,x_{\mathrm{clip}} + 2(K-i)\,x_{\mathrm{b}}}{K}, i = \dfrac{K}{2}+1,...,K \end{cases} . \tag{3.1.1.4}$$

Оба региона, CGR који покрива $[-x_{\mathrm{b}}, x_{\mathrm{b}}]$ и PGR који покрива $[-x_{\mathrm{clip}}, -x_{\mathrm{b}}) \cup (x_{\mathrm{b}}, x_{\mathrm{clip}}]$ су симетрично подељени на K униформних ћелија (видети Слику 3.1.1.2. где је приказана преносна карактеристика PWUQ за N = 16).



Слика 3.1.1.2. Преносна карактеристика PWUQ за $N = 2K$ = 16.

Због симетрије, сваки од региона [0, x_b] и (x_b, x_{clip}] подељен је на $K/2$ униформних ћелија. Пошто ова два униформна квантизера чине наш нови модел PWUQ, нотација нивоа одлучивања у CGR се завршава индексом $K/2$, док се индекс нивоа одлучивања у PGR повећава до K.

Даље дефинишемо величине корака квантизације Δ^{CGR} и Δ^{PGR} као униформне ширине ћелија $\Re_i$ у $\mathfrak{R}^{CGR}$ и $\mathfrak{R}^{PGR}$, респективно:

$$\Delta^{CGR} = \frac{2x_b}{K} = \frac{2\psi x_{clip}}{K}, \tag{3.1.1.5}$$

$$\Delta^{PGR} = \frac{2\left(x_{clip} - x_b\right)}{K} = \frac{2\left(1-\psi\right)x_{clip}}{K}. \tag{3.1.1.6}$$

Уводимо параметар ψ као однос прага границе и прага одсецања $\psi = x_b/x_{clip}$, за који, а у складу са дефиницијом PWUQ модела који описујемо (видети Слику 3.1.1.1.), важи $\psi < 1$ или прецизније $\psi < 0.5$. Посматрајући област испод Лапласове функције густине вероватноће, одлучујемо се на повећање ширине квантизационе ћелије у $\mathfrak{R}^{PGR}$ док се истовремено и последично смањује величина ширине квантизационе ћелије у $\mathfrak{R}^{CGR}$.

Описани симетрични PWUQ мапира реалну вредност $x \in \mathbb{R}$ у један од репрезентационих нивоа y_i

$$y_{-i} = -y_i, i = 1, 2, ..., K, \tag{3.1.1.7}$$

одређених као средишње тачке одговарајућих квантизационих ћелија $\Re_i \in \mathfrak{R}^{CGR}$ за $i = 1, ..., K/2$ и $\Re_i \in \mathfrak{R}^{PGR}$ где је $i = K/2+1, ..., K$

$$y_i = \begin{cases} \left(i - \frac{1}{2}\right)\Delta^{CGR}, i = 1, ..., \frac{K}{2} \\ x_b + \left(i - \frac{K+1}{2}\right)\Delta^{PGR}, i = \frac{K}{2}+1, ..., K \end{cases}. \tag{3.1.1.8}$$

Како би се сагледале перформансе описаног PWUQ за дату битску брзину r ($r = \log_2 N$), потребно је одредити грануларну дисторзију која потиче од квантизације у регионима $\Re^{\mathrm{CGR}}$ и $\Re^{\mathrm{PGR}}$, означене као $D_{\mathrm{g}}^{\mathrm{CGR}}$ и $D_{\mathrm{g}}^{\mathrm{PGR}}$

$$D_{\mathrm{g}}^{\mathrm{PWUQ}} = D_{\mathrm{g}}^{\mathrm{CGR}} + D_{\mathrm{g}}^{\mathrm{PGR}}, \tag{3.1.1.9}$$

$$D_{\mathrm{g}}^{\mathrm{PWUQ}} = \frac{1}{12}\left(\left(\Delta^{\mathrm{CGR}}\right)^2 P^{\mathrm{CGR}} + \left(\Delta^{\mathrm{PGR}}\right)^2 P^{\mathrm{PGR}}\right). \tag{3.1.1.10}$$

P^{CGR} и P^{PGR} представљају вероватноће да улазни одмерак x припада $\Re^{\mathrm{CGR}}$ и $\Re^{\mathrm{PGR}}$, респективно:

$$P^{\mathrm{CGR}} = 2\int_0^{x_{\mathrm{b}}} p(x)\,dx, \tag{3.1.1.11}$$

$$P^{\mathrm{PGR}} = 2\int_{x_{\mathrm{b}}}^{x_{\mathrm{clip}}} p(x)\,dx. \tag{3.1.1.12}$$

Подсетимо да средња вредност представља меру централне тенденције и специфицира где реална вредност x тежи да формира кластер, док стандардна девијација σ, показује како се крива шири од селектоване средње вредности и формира меру дисперзије [2]. Такође, познато је да се у неограниченом амплитудском домену, функција кумулативне расподеле (*CDF*), означена са F_{CDF}, може одредити као:

$$F_{\mathrm{CDF}}(b) = \int_{-\infty}^{b} p(x)dx, \tag{3.1.1.13}$$

где CDF задовољава услов $F_{\mathrm{CDF}}(\infty) = 1$.

За симетричне функције густине вероватноће, као што је дато у (3.1.1.2) важи:

$$\Phi(-b,b) = F_{\mathrm{CDF}}(b) - F_{\mathrm{CDF}}(-b) = 2\int_0^{b} p(x)\,dx, \tag{3.1.1.14}$$

где $\Phi(-b,b)$ представља вероватноћу да реална вредност одмерака x окарактерисана са $p(x)$ припада интервалу $[-b, b]$.

Позивајући се на (3.1.1.13) и (3.1.1.14), за P^{CGR} и P^{PGR}, налазимо:

$$P^{\mathrm{CGR}} = \Phi\left(-x_{\mathrm{b}}, x_{\mathrm{b}}\right) = 2\int_{0}^{x_{\mathrm{b}}} p\left(x\right)dx = 1-\exp\left\{-\sqrt{2}x_{\mathrm{b}}\right\}, \tag{3.1.1.15}$$

$$P^{\mathrm{PGR}} = \Phi\left(-x_{\mathrm{clip}}, x_{\mathrm{clip}}\right) - \Phi\left(-x_{\mathrm{b}}, x_{\mathrm{b}}\right) = 2\int_{x_{\mathrm{b}}}^{x_{\mathrm{clip}}} p\left(x\right)dx = \exp\left\{-\sqrt{2}x_{\mathrm{b}}\right\} - \exp\left\{-\sqrt{2}x_{\mathrm{clip}}\right\}. \tag{3.1.1.16}$$

Заменом (3.1.1.5), (3.1.1.6), (3.1.1.15) и (3.1.1.16) у (3.1.1.10) следи:

$$D_g^{\mathrm{PWUQ}} = C\left[\psi^2\left(1-\exp\left\{-\sqrt{2}\psi\, x_{\mathrm{clip}}\right\}\right) + \left(1-\psi\right)^2\left(\exp\left\{-\sqrt{2}\psi\, x_{\mathrm{clip}}\right\} - \exp\left\{-\sqrt{2}x_{\mathrm{clip}}\right\}\right)\right], \tag{3.1.1.17}$$

где је $C = 4\ x_{\mathrm{clip}}^2/(3N^2)$. Даљим одређивањем првог извода $D_{\mathrm{g}}^{\mathrm{PWUQ}}$ по ψ и изједначавањем са нулом

$$\frac{\partial D_g^{\mathrm{PWUQ}}}{\partial \psi} = 0, \tag{3.1.1.18}$$

добија се:

$$x_b = \psi\, x_{clip} = \frac{1}{\sqrt{2}}\ln\left[\frac{1+\frac{x_{clip}}{\sqrt{2}} - \sqrt{2}\psi\, x_{clip}}{\psi + \left(1-\psi\right)\exp\left\{-\sqrt{2}x_{clip}\right\}}\right]. \tag{3.1.1.19}$$

Из израза (3.1.1.19), а на основу дискусије о одређивању x_{b} или ψ, очигледна је потреба за применом итеративног нумеричког метода:

$$\psi^{(i)} = \frac{1}{\sqrt{2}x_{\mathrm{clip}}}\ln\left(\frac{1+\frac{x_{\mathrm{clip}}}{\sqrt{2}}\left(1-2\psi^{(i-1)}\right)}{\psi^{(i-1)} + \left(1-\psi^{(i-1)}\right)\exp\left\{-\sqrt{2}x_{\mathrm{clip}}\right\}}\right). \tag{3.1.1.20}$$

Проналажењем другог извода грануларне дисторзије дате у (3.1.1.17) по ψ налазимо:

$$\frac{\partial^2 D_g^{\mathrm{PWUQ}}}{\partial \psi^2} = 1-\exp\left\{-\sqrt{2}x_{\mathrm{clip}}\right\} + x_{\mathrm{clip}}\exp\left\{-\sqrt{2}\psi\, x_{\mathrm{clip}}\right\}\left[2\sqrt{2} + x_{\mathrm{clip}}\left(1-2\psi\right)\right]. \tag{3.1.1.21}$$

Како важи $x_b \le x_{clip}/2$, $2\sqrt{2} + x_{clip}(1-2\psi) > 2\sqrt{2}$, може се закључити да је D_g^{PWUQ} конвексна функција од ψ за праг централног грануларног региона x_b, где је $x_b \in (0, x_{clip}/2]$. Прецизније, за $\partial^2 D_g^{PWUQ}/\partial\psi^2 > 0$, D_g^{PWUQ} је конвексна функција прага централног грануларног региона x_b тако да постоји јединствено x_b^* за које је D_g^{PWUQ} минимално. Иницијализација је у раду [17] вршена са $\psi(0) = 0.5$ обзиром да се тада PWUQ модел подудара са униформним квантизером. Дефинишемо критеријум заустављања итеративног алгоритма, који је задовољен када апсолутна грешка у израчунавању ψ

$$\varepsilon^{(i)} = \left|\psi^* - \psi^{(i-1)}\right|, \tag{3.1.1.22}$$

мања од od $\varepsilon_{min} = 10^{-4}$.

Алгоритам 3.1.1.1. - PWUQ_Laplacian (N, x_{clip}, ε_{min}) – итеративна параметризација PWUQ за Лапласову функцију густине вер. и дати x_{clip} – одређивање ψ^* и Φ ($-x_b^*$, x_b^*)

Улаз: Укупан број квантизационих нивоа N, предефинисана вредност x_{clip}, $\varepsilon_{min} << 1$
Излаз 1: ψ^*, $\Phi(-x_b^*, x_b^*)$
Излаз 2: Δ^{CGR}, Δ^{PGR}, $\{y_{-i}, y_i\}$, $\{x_{-i}, x_i\}$, $i = 1,2,..., K$

Иницијализација $i \leftarrow 0$,
$\psi^* \leftarrow 0.5$,
$\varepsilon(0) \leftarrow 0.5$
while $\varepsilon(i) > \varepsilon_{min}$ do
    $i \leftarrow i+1$
    $\psi(i-1) \leftarrow \psi^*$
    срачунати $\psi(i)$ употребом (3.1.1.20)
    $\psi^* \leftarrow \psi(i)$
    срачунати $\varepsilon(i)$ употребом (3.1.1.22)
end while
$x_b^* \leftarrow \psi^* \times x_{clip}$
срачунати $\Phi(-x_b^*, x_b^*)$ употребом (3.1.1.15)
вратити ψ^*, $\Phi(-x_b^*, x_b^*)$
срачунати Δ^{CGR}, Δ^{PGR}, $\{y_{-i}, y_i\}$, $\{x_{-i}, x_i\}$, $i = 1, 2, ..., K$

Позивајући се на (3.1.1.11), (3.1.1.12) и (3.1.1.13) срачунавамо даље λ, као вероватноћу да одмерак из неограниченог амплитудског домена x, са pdf $p(x)$, припада грануларном региону:

$$\lambda = \frac{P^{\mathrm{CGR}} + P^{\mathrm{PGR}}}{F_{\mathrm{CDF}}(\infty)} = P^{\mathrm{CGR}} + P^{\mathrm{PGR}}. \tag{3.1.1.23}$$

Такође можемо нагласити да x_{clip}, x_{b} и N директно утичу на дисторзију. Ако вредност прага одсецања x_{clip} има значајно малу вредност, тачност квантизације се може снизити или деградирати јер ће се превише одмерака одрезати (одбацити) [12], [58]. Нагласимо да ефекат одсецања нулира дисторзију прекорачења, тако да се додатно може утицати на свеукупну дисторзију ако праг одсецања није подобро одабран.

3.1.2 Анализа експерименталних и нумеричких резултата

Управо одређивање прага одсецања представља круцијални задатак за остваривање најбољих перформанси датог квантизационог задатка. Можемо предвидети да поред прага одсецања x_{clip}, и праг разграничења грануларних области, x_{b}, такође афектира грануларну дисторзију, јер се CGR и PGR дисторзија, које заједно одређују укупну дисторзију, опозитно понашају у односу на вредност прага разграничења грануларних области x_{b}. Наиме, за фиксни и једнаки број квантизационих нивоа у $\Re^{\text{CGR}}$ и $\Re^{\text{PGR}}$, са смањењем вредности прага разграничења грануларних облати, CGR дисторзија се редукује уз истовремено повећање PGR дисторзије. Тако смањење ширине $\Re^{\text{CGR}}$ може условити значајну редукцију дисторзије у CGR, што истовремено и последично резултира нежељеним, али очекиваним порастом дисторзије у PGR, која је примарно намењена за репове Лапласове функције густине вероватноће. Из свега наведеног може се закључити да је фино подешавање прага разграничења грануларних области $\pm x_{\text{b}}$ и прага одсецања $\pm x_{\text{clip}}$, кључни или доминантно одлучујући оптимизациони фактор када се разматра Лапласова функција густине вероватноће са тешким реповима.

У наставку показујемо да за Лапласову функцију густине вероватноће, описани PWUQ показује значајно боље перформансе у односу на униформни квантизер, при чему се као мера поређења користи SQNR:

$$\text{SQNR}^{\text{PWUQ}}[\text{dB}] = -10 \cdot \log_{10}\left(D_g{}^{\text{PWUQ}}\right). \tag{3.1.2.1}$$

У поређењу са униформни квантизер за исте битске брзине:

$$\text{SQNR}^{\text{UQ}}[\text{dB}] = -10 \cdot \log_{10}\left(D_g{}^{\text{UQ}}\right), \tag{3.1.2.2}$$

$$D_g{}^{\text{UQ}} = \frac{x_{\text{UQ}}{}^2}{3N^2}, \tag{3.1.2.3}$$

интересантно је приметити да уколико претпоставимо исте ширине грануларних региона и вредности прагова одсецања за PWUQ и униформни квантизер, тј. $x_{\text{clip}} = x_{\text{UQ}}$, долази се до крајње формуле у затвореном облику, која даје детаљан увид у

перформансни добитак, тј. SQNR добитак, који описани PWUQ показује у односу на униформни квантизер:

$$\delta = -10\log\left(\frac{D_g^{\text{PWUQ}}}{D_g^{\text{UQ}}}\right) = -10\log\left(4\left[\psi^2\left(P^{\text{CGR}} + P^{\text{PGR}}\right) + (1-2\psi)P^{\text{PGR}}\right]\right). \qquad (3.1.2.4)$$

Очигледно, за $\psi = 0.5$ и исти број нивоа N, PWUQ и униформни квантизер се подударају тако да је $\text{SQNR}^{\text{PWUQ}}$ заправо једнак SQNR^{UQ}. Најзначајнији корак у дизајнирању описаног PWUQ је управо одређивање кључног параметра ψ. За дати праг одсецања x_{clip}, може се срачунати ψ^* коришћењем итеративног процеса специфицираног са (3.1.1.20). Затим се може одредити $x_b^* = \psi^* \cdot x_{\text{clip}}$ и сви остали параметри описаног PWUQ, као и перформансни индикатор $\text{SQNR}^{\text{PWUQ}}$. Како је предложена Лапласова функција густине вероватноће са дугим реповима неограничена, може се прихватити да се већина одмерака функције густине вероватноће налази управо у грануларном региону тако да је $\lambda^{\text{GR}} = P^{\text{CGR}} + P^{\text{PGR}} = 0.9999$. Такође се може анализирати и случај када су праг одсецања као и праг разграничења грануларних области, $x_{\text{clip}}^{\text{GR}}$ и x_b^{GR}, као и њихов међусобни однос ψ^{GR}, одређени из услова да је $\lambda^{\text{GR}} = 0.9999$. Овде се GR нотација користи за грануларни регион. Полазећи од (3.1.1.15) и (3.1.1.16) за $\lambda^{\text{GR}} = 0.9999$, налазимо:

$$x_{\text{clip}}^{\text{GR}} = -\frac{1}{\sqrt{2}}\ln\left(1-\lambda^{\text{GR}}\right) = 2\sqrt{2}\ln(10). \qquad (3.1.2.5)$$

За r у опсегу од 5 до 8 бит/одмерак, могу се срачунати ψ^*, $x_b^* = \psi^* \cdot x_{\text{clip}}$, $\Phi(-x_b^*, x_b^*)$, λ^*, SQNR^{UQ}, $\text{SQNR}^{\text{PWUQ}}$ и δ, за различите вредности x_{clip} (видети таб. 3.1.2.1. и. 3.1.2.2.).

Наиме, упоредо са одређивањем $x_{\text{clip}}^{\text{GR}}$, датим у (3.1.2.5), могу се одредити и сетови параметара за x_{clip} за вредности дате у [57] и [2], које су предложили аутори *Hui* и *Jayant*, респективно. У циљу прецизнијег специфицирања ова три случаја, у раду [17] уведене се нотације [J], [H] и [GR] (или само GR), које се наводе у истој линији у суперскрипту (*superscript*).

Табела 3.1.2.1. Упоредни приказ главних параметара дизајна униформног квантизера и PWUQ (1. део).

r [бит/одмерак]		x_{clip}	ψ^*	λ^*	x_b*	Φ ($-x_b$*, x_b*)
5	[J]	4.4800	0.3100	0.9982	1.38880	0.8597
	[H]	4.9013	0.2996	0.9990	1.46843	0.8747
	GR	6.5127	0.2674	0.9999	1.74150	0.9148
6	[J]	5.3024	0.2903	0.9994	1.53929	0.8866
	[H]	5.8815	0.2789	0.9998	1.64035	0.9017
	GR	6.5127	0.2674	0.9999	1.74150	0.9148
7	[J]	6.1504	0.2740	0.9998	1.6852	0.9077
	[H]	6.8618	0.2616	0.9999	1.79505	0.9210
	GR	6.5127	0.2674	0.9999	1.74150	0.9148
8	[J]	7.0272	0.2589	0.9999	1.81934	0.9237
	[H]	7.8421	0.2467	0.9999	1.83465	0.9552
	GR	6.5127	0.2674	0.9999	1.74150	0.9148

Као што се може уочити на слици 3.1.2.2. или једноставним прорачуном из (3.1.1.4) и (3.1.1.5) ($\Delta^{CGR} / \Delta^{PGR} = \psi^* /(1-\psi^*)$) или из нумеричких резултата датих у првом реду табеле 3.1.2.1. за уочени случај важи да $\Delta^{PGR} \approx 2\Delta^{CGR}$, тако да $K\times(\Delta^{PGR} + \Delta^{CGR}) = 2K\times \Delta_{UQ}$ узрокује $\Delta_{UQ} \approx 1.5\ \Delta^{CGR}$ и $\Delta_{UQ} \approx 0.75\ \Delta^{PGR}$. Према томе, може се показати да у поређењу са униформним квантизером, прецизнија квантизација је извршена у $\Re^{CGR}$ коме припада већина фино квантованих одмерака претпостављене Лапласове функције густине вероватноће, док се у $\Re^{PGR}$, где су доминантно репови функције густине вероватноће, униформни квантизер постиже незнатно боље перформансе, а у складу са условом да је $\Delta_{UQ} < \Delta^{PGR}$. Наиме, имајући на уму сам облик Лапласове функције густине вероватноће, показује се оправданим тежња да се финије квантују одмерци функције густине вероватноће концентрисани око средње вредности, а који припадају $\Re^{CGR}$.

Табела 3.1.2.2. Упоредни приказ главних параметара дизајна униформног квантизера и PWUQ (2. део).

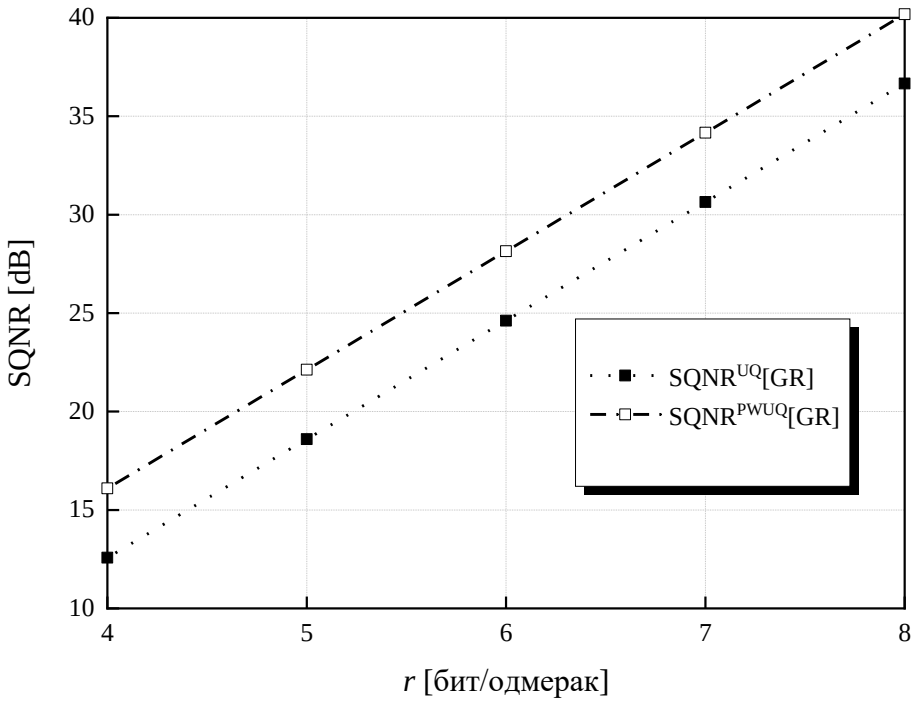
r [бит/одмерак]		x_{clip}	ψ^*	$SQNR^{UQ}$	$SQNR^{PWUQ}$	δ
5	[J]	4.4800	0.3100	21.8487	24.1089	2.2602
	[H]	4.9013	0.2996	21.0680	23.6010	2.5330
	GR	6.5127	0.2674	18.5990	22.1220	3.5230
6	[J]	5.3024	0.2903	26.4054	29.1938	2.7884
	[H]	5.8815	0.2789	25.5050	28.6522	3.1472
	GR	6.5127	0.2674	24.6196	28.1426	3.5230
7	[J]	6.1504	0.2740	31.1373	34.4466	3.3093
	[H]	6.8618	0.2616	30.1866	33.9106	3.7240
	GR	6.5127	0.2674	30.6402	34.1632	3.5230
8	[J]	7.0272	0.2589	36.0004	39.8178	3.8174
	[H]	7.8421	0.2467	35.0474	39.3096	4.2622
	GR	6.5127	0.2674	36.6608	40.1838	3.5230

Како је претпостављено $\psi^* < 0.5$, из $\Delta^{CGR} / \Delta^{PGR} = \psi^* / (1-\psi^*)$ се може наћи да је $\Delta^{CGR} < \Delta^{PGR}$, чиме се потврђује ужи $\Re^{CGR}$ коме припада већина одмерака Лапласове функције густине вероватноће, као и мања грешка квантизације. Као резиме, а полазећи од облика Лапласове функције густине вероватноће и начина спровођења параметризације описаног PWUQ, свеукупни SQNR добитак који PWUQ постиже у односу на униформни квантизер се може проценити и прецизно срачунати, што додатно потврђује коректност полазних поставки у моделирању новог PWUQ.

Ако се размотри x_{clip} срачунат из (3.1.2.5) за $\lambda^{GR} = 0.9999$ и за $\varepsilon < 10^{-4}$, ψ^{GR} постиже се вредност 0.2674 која је константна и независна од битске брзине. Наиме, могу се одредити и следеће фиксне вредности: $x_{clip} = x_{clip}{}^{GR} = 6.5127$, $\psi^{GR} = 0.2674$, $x_b{}^{GR} = 1.74150$ и $\Phi(-x_b{}^{GR}, x_b{}^{GR}) = 0.9148$, док $SQNR^{PWUQ}$ стриктно зависи од r. Интересантно је запажање да иако $SQNR^{PWUQ}$ зависи од r, SQNR добитак који PWUQ

остварује у односу на униформни квантизер, у тзв. GR случају, је такође константан и износи фиксно 3.523 dB (видети слику 3.1.2.1. за GR случај).

Како би пецизирали SQNR добитак у раду [17] смо кренули од формуле (3.1.2.4), из које тривијално следе закључци за фиксно срачунате ψ^{GR}, $x_{clip}{}^{GR}$ и $x_b{}^{GR}$ вредности. Коначно, може се истаћи да за $r = 8$ бит/одмерак, $SQNR^{PWUQ}[GR]$, одређен за $x_{clip} = x_{clip}{}^{GR} = 6.5127$, је већи у односу на $SQNR^{PWUQ}[H]$ [57] и $SQNR^{PWUQ}[J]$ [2], што се може објаснити погодним ефектом одсецања у GR случају [17], а у складу са претпоставком да је $\lambda^{GR} = 0.9999$, као и са чињеницом да се за разлику од [57] и [2], оптимизација амплитуде максималног оптерећења, овде наведене као праг одсецања, ради уз анулирање дисторзије прекорачења због примене самог ефекта одсецања.

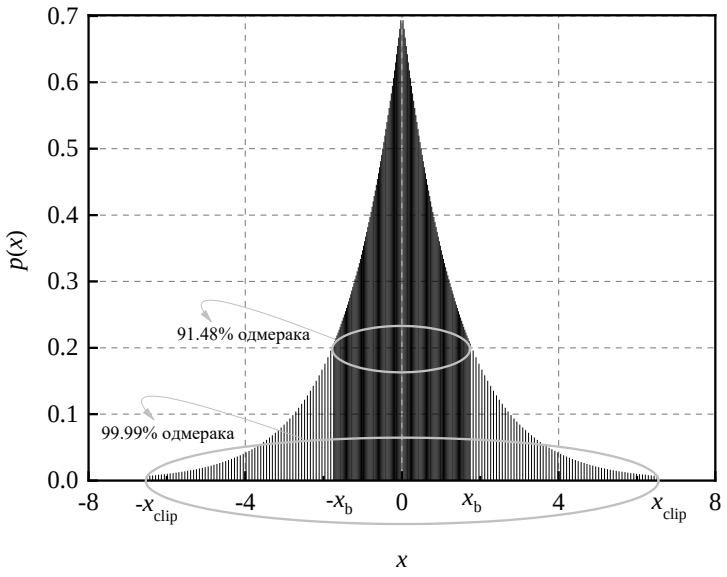


Слика 3.1.2.1. Константан SQNR добитак од 3.523 dB који PWUQ остварује у односу на униформни квантизер и GR за r од 4 бит/одмерак до 8 бит/одмерак.

Значајно је приметити, а у складу са табелом 3.1.2.1, да Лапласова функција густине вероватноће очекивано припада грануларном региону са $\lambda^{*[J]} = 0.9982$, $\lambda^{*[H]} = 0.9990$ за $r = 5$ бит/одмерак, док су за $r = 8$ бит/одмерак $\lambda^{*[J]}$ и $\lambda^{*[H]}$ једнаки $\lambda^{GR} = 0.9999$. Срачунавајући λ^* и $\Phi(-x_b^*, x_b^*)$ за наведена три случаја, може се потврдити чињеница да се подробно одабраним прагом одсецања, само мали део одмерака налази у грануларном региону, а близу прага одсецања или ван грануларног региона.

Такође, пажљивом опсервацијом табеле 3.1.2.1. може се приметити да је вероватноћа да амплитуда реалног омерака x припада $\Re^{CGR}$, расте са порастом битске брзине, иако и саме вредности $x_{clip}^{[J]}$ и $x_{clip}^{[H]}$ такође очекивано расту са порастом битске брзине. Такође, може се посебно истаћи и доминација коју $\Phi(-x_b^{GR}, x_b^{GR})$ показује у односу на $\Phi(-x_b^{*[J]}, x_b^{*[J]})$ и $\Phi(-x_b^{*[H]}, x_b^{*[H]})$ за исти опсег битских брзина и то за 5 бит/одмерак и 6 бит/одмерак, док се $\Phi(-x_b^{GR}, x_b^{GR})$ приближава ка $\Phi(-x_b^{*[J]}, x_b^{*[J]})$ и $\Phi(-x_b^{*[H]}, x_b^{*[H]})$ за $r = 7$ бит/одмерак и $r = 8$ бит/одмерак. Како су вредности $\Phi(-x_b^{GR}, x_b^{GR})$ и x_{clip}^{GR} константне, за $r = 7$ бит/одмерак је $x_{clip}^{GR} < x_{clip}^{[H]}$, што резултира $SQNR^{PWUQ}[GR] > SQNR^{PWUQ}[H]$. Слично, за $r = 8$ бит/одмерак и за $x_{clip}^{GR} < x_{clip}^{[J]}$ је $SQNR^{PWUQ}[GR] > SQNR^{PWUQ}[J]$. За највећу посматрану битску брзину, $r = 8$ бит/одмерак, поред вредности кључних параметара дизајна и процене перформанси униформног квантизера и PWUQ, у последњој врсти табеле 3.1.2.1. је дат приказ за GR случај, док су додатни дескриптивни показатељи дати на слици 3.1.2.2.

Са слике 3.1.2.2. се може приметити да 91.48 % одмерака Лапласове функције густине вероватноће припада $\Re^{CGR}$, док се 99.99 % сумарно налази у обе грануларне регије $\Re^{CGR} \cup \Re^{PGR}$, што указује да је одабрана вредност $\psi = \psi^{GR}$ таква да велики проценат одмерака припада $\Re^{CGR}$. У случају са $r = 8$ бит/одмерак одредили смо најмање вредности ψ^* у [J] и [H] случајевима, односно утврдили смо највеће апсолутне разлике од почетне вредности $\psi(0) = 0.5$. У ова два случаја, број итерација за одређивање ψ^* креће се до 25, при чему се вредности ψ^* поклапају са резултатима нумеричке оптимизације дисторзије по ψ, што значи да описани алгоритам брзо конвергира.

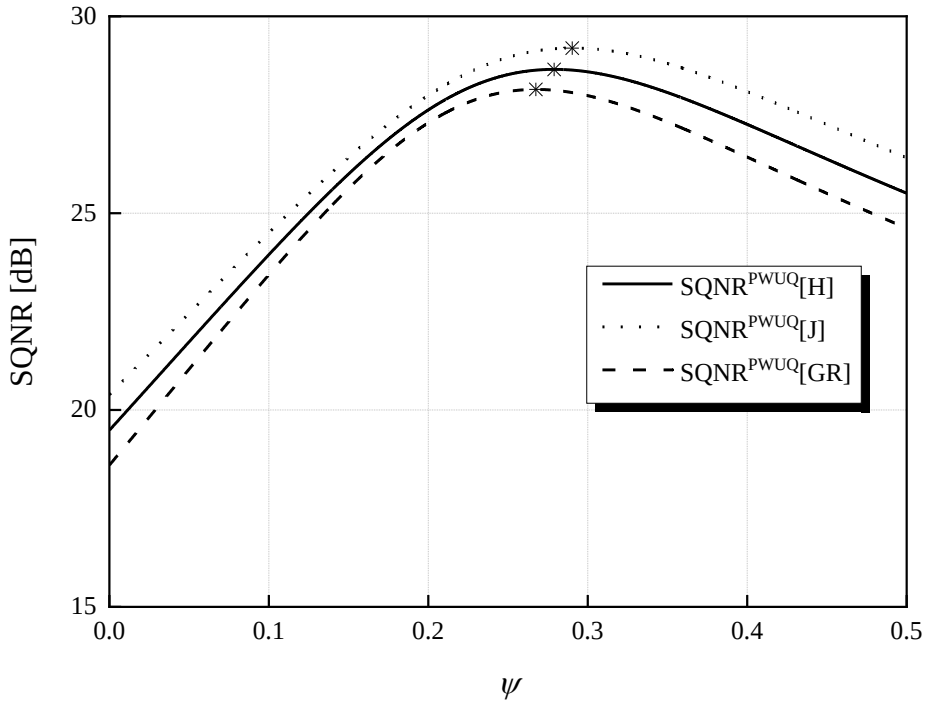


Слика 3.1.2.2. Припадност одмерака $\Re^{CGR}$ и $\Re^{PGR}$ за Лапласов извор и за GR случај (r = 8 бит/одмерак).

Како би илустровали значај и важност одређивања вредности ψ, али такође и подробност избора x_{clip}, слика 3.1.2.3. приказује зависност $SQNR^{PWUQ}$ од ψ за $r = 6$ бит/одмерак и за сва три случаја од интереса, за унапред одабране вредности x_{clip}^{GR}, $x_{clip}^{[H]}$, $x_{clip}^{[J]}$. Истакнимо да су вредности параметра ψ, које су резултати описаног итеративног алгоритма за разматрана три случаја, означени звездицама на слици 3.1.2.3. заиста оптималне и дају одговарајући максимум $SQNR^{PWUQ}$ за ова три случаја.

Да бисмо изабрали једну од вредности за x_{clip}, за дату брзину преноса r, можемо изабрати ону која даје највећи SQNR [17].

Коначно, као у [17], може се недвосмислено закључити да иако је опсег вредности ψ релативно узак, погодан одабир његове вредности је битан, обзиром да се управо његовим неповољним одабиром могу значајно деградирати перформансе описаног PWUQ. Ова опсервација додатно потврђује важност спроведене анализе и приложених резултата.



Слика 3.1.2.3. $SQNR^{PWUQ}$ зависност од ψ за [J], [H] и [GR] случај (r = 6 бит/одмерак).

Да резимирамо, оригиналност овог модела у области квантизације у обради сигнала огледа се у следећем:

- имајући у виду облик Лапласове функције густине вероватноће, предложена је нова идеја за поделу амплитудског опсега квантизера на два региона, CGR и PGR;

- за дате вредности x_{clip}, ширине ова два региона су оптимизоване коришћењем итеративног алгоритма, који задовољава услов минималне дисторзије PWUQ модела;

- коришћен је најједноставнији модел униформног квантизера, који се за једнаке битске брзине примењује у сваком од два региона, што дизајн модела који смо описали у овом одељку чини много једноставнијим у поређењу са многим неуниформним моделима квантизера доступним у литератури.

- Постигнут је значајан SQNR добитак у односу на униформни квантизер, што даље оправдава смисленост описаног модела.

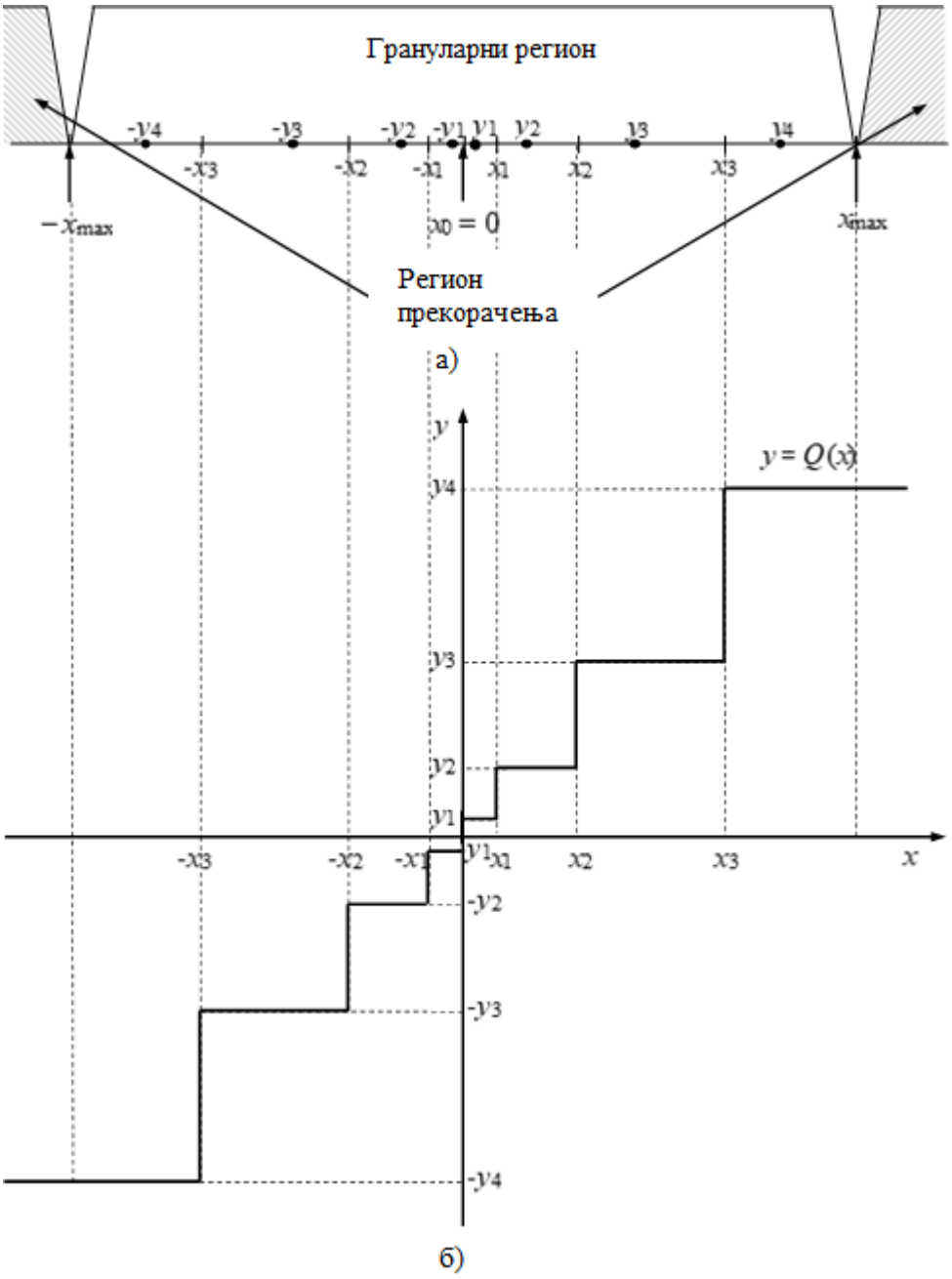
3.2 Параметризација симетричног квантиле квантизера за Лапласов извор

Бројни радови су дали општи оквир за подстицање дискуције о могућим тумачењима утицаја броја нивоа квантизације и ширине грануларног региона на грануларну дисторзију и дисторзију прекорачења [12] - [14], [30], [57], [58]. Имајући у виду дефиницију грануларног региона, као области у којој је дисторзија мала или на одговарајући начин ограничена [2], [57], главни компромис у дизајну квантизера представља прилагођавање амплитудској динамици сигнала уз минимизирање укупне дисторзије. С тим у вези, као што је приказано у [2], амплитуда максималног оптерећења је вероватно најзначајнији параметар квантизера.

У овом одељку се бавимо управо одређивањем грануларног региона симетричног квантиле квантизера (SQQ) који описујемо у [18], као и његовим прилагођењем како би SQQ модел квалификовали са побољшаном укупном ефикасношћу квантизације за улазе са Лапласовом функције густине вероватноће. Важан аспект наше дискусије дизајна SQQ базира се на процени и анализи SQNR. Конкретно, постављајући за задатак оптимизацију у броју пренетих битова, на основу процене средње-квадратне грешке дисторзије и налажења њеног минимума, директно израчунавамо амплитуда максималног оптерећења, што нам омогућава да вршимо даље анализе и изводимо закључке.

3.2.1 Пројектовање симетричног квантиле квантизера

Одмерци улазног сигнала или уопштено улаза, које је потребно квантовати, се додељују неком од два региона – грануларном (унутрашњем) региону и региону прекорачења (спољашњем), који су за симетричне квантизере одвојени симетрично постављеним амплитудама максималног оптерећења означеним као $-x_{max}$ и x_{max}, респективно [2] (видети слику 3.2.1.1.).



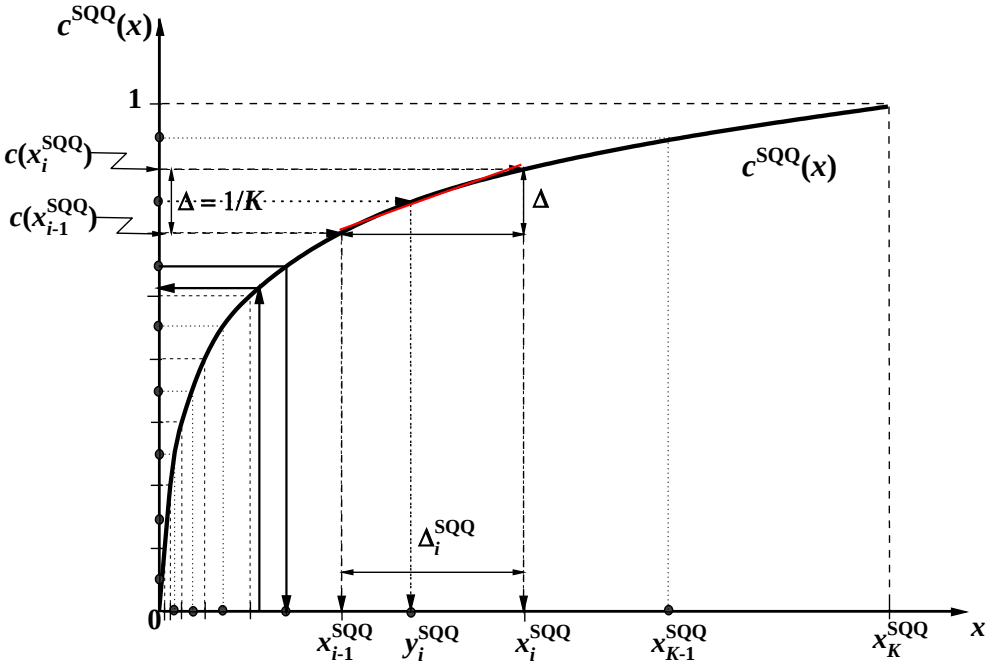
Слика 3.2.1.1. (а) Грануларни и регион прекорачења симетричног неуниформног квантизера са $N = 8$ квантизационих нивоа.

(б) Карактеристика неуниформног квантизера $Q(x)$ за $N = 8$.

Претпостављамо Лапласову функцију густине вероватноће нулте средње вредности и јединичне варијансе:

$$p(x)=\frac{1}{\sqrt{2}}\exp\left\{-\sqrt{2}|x|\right\}. \qquad (3.2.1.1)$$

Позивајући се на лог-конкавност и доказану конвекност MSE дисторзије за Лапласову функцију густине вероватноће [29], у овом одељку се детаљно описује одређивање параметара SQQ. Модел SQQ предложен у [20], има N нивоа и дефинисан је пресликавањем $Q:\mathbb{R}\to Y$ [2], где је $Q(\cdot)$ симетрична карактеристика квантизера, $\mathbb{R}$ је скуп реалних бројева, $Y \equiv \{y_{-N/2},\ldots, y_{-1}, y_1, \ldots, y_{N/2}\} \subset \mathbb{R}$ скуп репрезентационих нивоа (видети сл. 3.2.1.1.), који чини кодну књигу величине $|Y| = N = 2K$. Код симетричних квантизера са парним бројем непреклапајућих квантизационих ћелија, квантизациони нивои су симетрично распоређени око средње вредности.



Слика 3.2.1.2. Позитивни део непарно симетричне компресорске функције $c^{SQQ}(x)$.

Без умањења општости, усмерићемо пажњу само на K позитивних квантизационих ћелија, као што је приказано на сл.3.2.1.2. на којој је представљен и позитивни део непарно симетричне компресорске функције $c^{\mathrm{SQQ}}(x)$.

Сет потребних параметара који у потпуности описују SQQ са N-нивоа је [20]:

{функција густине квантизационих нивоа; компресорска функција; нивои одлучивања; репрезентациони нивои; амплитуда максималног оптерећења; дисторзија}.

Функција густине квантизационих нивоа овог модела SQQ ја дата као:

$$\lambda(x) = p^{1/3}(x)\left[2\int_0^{x_K^{\mathrm{SQQ}}}\left[p(t)\right]^{1/3}dt\right]^{-1}. \qquad (3.2.1.2)$$

Даљом заменом (3.2.1.1) у (3.2.1.2), израз за функцију густине квантизационих нивоа постаје:

$$\lambda(x) = \left[3\sqrt{2}\exp\left\{\frac{\sqrt{2}x}{3}\right\}\left(1-\exp\left\{-\frac{\sqrt{2}x_K^{\mathrm{SQQ}}}{3}\right\}\right)\right]^{-1}. \qquad (3.2.1.3)$$

Наша непарна компресорска функција је дефинисана са $c^{\mathrm{SQQ}}(x) : [0,+\infty) \rightarrow [0, 1)$

$$c^{\mathrm{SQQ}}(x) = 2\int_0^x \lambda(t)dt = \left(1-\exp\left\{-\frac{\sqrt{2}x}{3}\right\}\right)\left(1-\exp\left\{-\frac{\sqrt{2}x_K^{\mathrm{SQQ}}}{3}\right\}\right)^{-1},\ x \geq 0. \qquad (3.2.1.4)$$

Ненегативни нивои одлуке описаног SQQ одређени су компресорском функцијом:

$$x_i^{\mathrm{SQQ}} = c^{\mathrm{SQQ}\,-1}\left(\frac{i}{K}\right),\ i = 0,1,...,K-1, x_0^{\mathrm{SQQ}} = 0,\ x_K^{\mathrm{SQQ}} = +\infty, x_{-i}^{\mathrm{SQQ}} = -x_i^{\mathrm{SQQ}}, i = 1,2,...,K, \quad (3.1.1.5)$$

док се негативни парови нивоа одлуке директно одређују из услова симетрије. За разлику од компандинг модела квантизера, где су репрезентациони нивои такође одређени из компресорске функције, код SQQ сваки од репрезентационих нивоа је одређен из услова центроида одговарајуће квантизационе ћелије $\left[x_{i-1}^{\mathrm{SQQ}}, x_i^{\mathrm{SQQ}}\right)$ за претпостављену $p(x)$:

$$y_i^{\mathrm{SQQ}} = E_p\left\{X \middle| x_{i-1}^{\mathrm{SQQ}} \le X < x_i^{\mathrm{SQQ}}\right\}, y_{-i}^{\mathrm{SQQ}} = -y_i^{\mathrm{SQQ}}, i = 1,2,...,K\,, \tag{3.2.1.6}$$

где X и $E_p\{\cdot\}$ означавају континуалну случајну променљиву са претпостављеном $p(x)$ и очекивање за исту $p(x)$.

$$y_i^{\mathrm{SQQ}} = \int_{x_{i-1}^{\mathrm{SQQ}}}^{x_i^{\mathrm{SQQ}}} xp(x)dx \left[\int_{x_{i-1}^{\mathrm{SQQ}}}^{x_i^{\mathrm{SQQ}}} p(x)dx\right]^{-1}, i = 1,2,...,K\,, \tag{3.2.1.7}$$

$$y_K^{\mathrm{SQQ}} = x_{K-1}^{\mathrm{SQQ}} + \frac{1}{\sqrt{2}}\,. \tag{3.2.1.8}$$

За претпостављену функцију густине вероватноће и i = 1, 2, ..., K−1, изводимо:

$$y_i^{\mathrm{SQQ}} = x_{i-1}^{\mathrm{SQQ}} + \frac{1}{\sqrt{2}} - \frac{\Delta_i^{\mathrm{SQQ}} \exp\left\{-\sqrt{2}\Delta_i^{\mathrm{SQQ}}\right\}}{1-\exp\left\{-\sqrt{2}\Delta_i^{\mathrm{SQQ}}\right\}}, \tag{3.2.1.9}$$

$$\Delta_i^{\mathrm{SQQ}} = x_i^{\mathrm{SQQ}} - x_{i-1}^{\mathrm{SQQ}} = \frac{3}{\sqrt{2}} \ln\left(\frac{K-i+1}{K-i}\right). \tag{3.2.1.10}$$

Δ_i^{SQQ} је растојање између суседних нивоа одлуке за дати модел SQQ.

Увођењем $x_K^{\mathrm{SQQ}} = +\infty$ у изразима (3.2.1.3) и (3.2.1.4), израз за функцију густине квантизационих нивоа (3.2.1.11) као и израз за компресорску функцију (3.2.1.12) описаног модела SQQ постају :

$$\lambda(x) = \frac{1}{3\sqrt{2}} \exp\left\{-\frac{\sqrt{2}x}{3}\right\}, \tag{3.2.1.11}$$

$$c^{\mathrm{SQQ}}(x) = \left(1 - \exp\left\{-\frac{\sqrt{2}x}{3}\right\}\right), x \ge 0\,. \tag{3.2.1.12}$$

Прихватањем овако дефинисане $c^{\mathrm{SQQ}}(x)$, ненегативни нивои одлуке SQQ дефинисани изразом (3.2.1.5) се могу одредити као у (3.2.1.13):

$$x_i^{\mathrm{SQQ}} = \frac{3}{\sqrt{2}} \ln\left(\frac{K}{K-i}\right), i = 0,1,..., K-1\,, \tag{3.2.1.13}$$

и за $i = K-1$, из израза (3.2.1.5) и (3.2.1.13) директно изводимо формулу затвореног облика за праг одсецања:

$$x_{K-1}^{\mathrm{SQQ}} = \frac{3}{\sqrt{2}} \ln(K) . \tag{3.2.1.14}$$

Применом Бенетовог интеграла (*Bennett's integral*) [10] за апроксимацију грануларне дисторзије SQQ добија се:

$$D_{\mathrm{g}}^{\mathrm{SQQ}} = \frac{2\Delta^2}{12} \int_0^{x_{K-1}^{\mathrm{SQQ}}} \left[\frac{\partial c^{\mathrm{SQQ}}(x)}{\partial x} \right]^{-2} p(x) dx , \tag{3.2.1.15}$$

како за модел важи $\Delta = 2/N = 1/K$, изводи се израз за грануларну дисторзију:

$$D_{\mathrm{g}}^{\mathrm{SQQ}} = \frac{9}{2N^2} \left(1 - \exp\left\{ -\frac{\sqrt{2} x_{K-1}^{\mathrm{SQQ}}}{3} \right\} \right). \tag{3.2.1.16}$$

Даљом заменом (3.2.1.14) у (3.2.1.16) долази се до формуле затвореног облика за грануларну дисторзију:

$$D_{\mathrm{g}}^{\mathrm{SQQ}} = \frac{9}{2N^2} \left(1 - \frac{2}{N} \right). \tag{3.2.1.17}$$

Како би одредили и дисторзију прекорачења користимо y_K^{SQQ} срачунато изразом (3.2.1.8):

$$D_{\mathrm{over}}^{\mathrm{SQQ}} = 2 \int_{x_{K-1}^{\mathrm{SQQ}}}^{x_K^{\mathrm{SQQ}}} \left(x - y_K^{\mathrm{SQQ}} \right)^2 p(x) dx , \tag{3.2.1.18}$$

$$D_{\mathrm{over}}^{\mathrm{SQQ}} = \frac{\left(x_{K-1}^{\mathrm{SQQ}} + \frac{1}{\sqrt{2}} \right)^2 + \frac{1}{2} - y_K^{\mathrm{SQQ}} \left(2x_{K-1}^{\mathrm{SQQ}} + \sqrt{2} \right) + y_K^{\mathrm{SQQ}2}}{\exp\left\{ \sqrt{2} x_{K-1}^{\mathrm{SQQ}} \right\}} = \frac{1}{2} \exp\left\{ -\sqrt{2} x_{K-1}^{\mathrm{SQQ}} \right\}. \tag{3.2.1.19}$$

Директном заменом (3.2.1.14) у (3.2.1.19) долази се до крајње елегантног израза за дисторзију прекорачења:

$$D_{\mathrm{over}}^{\mathrm{SQQ}} = \frac{4}{N^3} . \tag{3.2.1.20}$$

Сумирањем дисторзија датих изразима (3.2.1.17) и (3.2.1.20) одређује се укупна дисторзија:

$$D_{\text{Asym}}^{\text{SQQ}} = \frac{9}{2N^2}\left(1-\frac{10}{9N}\right). \tag{3.2.1.21}$$

Asym у ознаци за дисторзију у изразу (3.2.1.21) односи се на примену Бенетовог интеграла у датој асимптотској анализи.

Како би утврдили тачност изведене асимптотске формуле за тоталну дисторзију, у наставку изводимо тачну формулу за тоталну дисторзију $D_{\text{Exact}}^{\text{SQQ}}$ под истим условима, тј. за Лапласову функцију густине вероватноће нулте средње вредности и јединичне варијансе.

За било коју парну вредност броја квантизационих нивоа N, тачна дисторзија се срачунава као:

$$D_{\text{Exact}}^{\text{SQQ}} = 2\sum_{i=1}^{K}\int_{x_{i-1}^{\text{SQQ}}}^{x_i^{\text{SQQ}}}\left(x-y_i^{\text{SQQ}}\right)^2 p(x)dx, \tag{3.2.1.22}$$

$$D_{\text{Exact}}^{\text{SQQ}} = 2\sum_{i=1}^{K}\int_{x_{i-1}^{\text{SQQ}}}^{x_i^{\text{SQQ}}} x^2 p(x)dx - 4\sum_{i=1}^{K}\int_{x_{i-1}^{\text{SQQ}}}^{x_i^{\text{SQQ}}} xy_i^{\text{SQQ}} p(x)dx + 2\sum_{i=1}^{K}\int_{x_{i-1}^{\text{SQQ}}}^{x_i^{\text{SQQ}}} y_i^{\text{SQQ2}} p(x)dx, \tag{3.2.1.23}$$

при чему је:

$$\int_{-\infty}^{+\infty} x^2 p(x)dx = 1, \tag{3.2.1.24}$$

те се дисторзија егзактно одређује као:

$$D_{\text{Exact}}^{\text{SQQ}} = 1 - 2\sum_{i=1}^{K-1}\left(y_i^{\text{SQQ}}\right)^2 P_i^{\text{SQQ}}, \tag{3.2.1.25}$$

$$P_i^{\mathrm{SQQ}} = \int_{x_{i-1}^{\mathrm{SQQ}}}^{x_i^{\mathrm{SQQ}}} p(x)\,dx = \frac{\exp\left\{-\sqrt{2}x_{i-1}^{\mathrm{SQQ}}\right\} - \exp\left\{-\sqrt{2}x_i^{\mathrm{SQQ}}\right\}}{2}, \tag{3.2.1.26}$$

Коришћењем (3.2.1.13) изводи се коначни израз као:

$$P_i^{\mathrm{SQQ}} = \frac{1}{2K^3}\left(3(K-i)^2 + 3(K-i) + 1\right), \tag{3.2.1.27}$$

$$D_{\mathrm{Exact}}^{\mathrm{SQQ}} = 1 - 2\sum_{i=1}^{K-1}\left(y_i^{\mathrm{SQQ}}\right)^2 \frac{1}{2K^3}\left(3(K-i)^2 + 3(K-i) + 1\right). \tag{3.2.1.28}$$

Дакле, може се закључити да је потребан скуп параметара SQQ са *N*-нивоа одређен као:

$$\left\{\begin{array}{l} \lambda(x) = \dfrac{1}{3\sqrt{2}}\exp\left\{-\dfrac{\sqrt{2}x}{3}\right\}; c^{\mathrm{SQQ}}(x) = \left(1 - \exp\left\{-\dfrac{\sqrt{2}x}{3}\right\}\right), x \geq 0; \\ x_i^{\mathrm{SQQ}} = \dfrac{3}{\sqrt{2}}\ln\left(\dfrac{K}{K-i}\right), i = 0, 1, \ldots, K-1; y_i^{\mathrm{SQQ}} = E_p\left\{X \,\middle|\, x_{i-1}^{\mathrm{SQQ}} \leq X < x_i^{\mathrm{SQQ}}\right\}, i = 1, 2, \ldots, K; \\ x_{K-1}^{\mathrm{SQQ}} = \dfrac{3}{\sqrt{2}}\ln(K); D_{\mathrm{Asym}}^{\mathrm{SQQ}} = \dfrac{9}{2N^2}\left(1 - \dfrac{10}{9N}\right). \end{array}\right\}.$$

(3.2.1.29)

Уколико су нивои одлуке и репрезентациони нивои адекватно одређени, укупна дисторзија је мања, што даље може условити смањење у броју битова потребних за постизање одређене дисторзије, која се сматра прихватљивом. Нагласимо да је SQQ, модификовани модел неуниформног квантизера тј. компандора или компандинг квантизера. Уопштено, компандинг квантизер, овде означен као SCQ (*Scalar Companding Quantizer),* је концептуално реализован каскадном везом три функционална блока: компресора, униформног квантизера и експандора [2]. Посматрани SQQ модел користи предност компандора за рачунање нивоа одлуке, док су репрезентациони нивои оптимално дефинисани, чиме се прави основна разлика у односу на компандорски модел где су и нивои одлучивања и репрезентациони нивои одређени у складу са специфичном функцијом компресора. Стога истичемо да се наши SQQ и SCQ модели одликују сличним нивоом сложености дизајна и имплементације,

тако да се оба модела могу користити за ефикасну иницијализацију *Lloyd-Max* алгоритма [59], [60]. Штавише, SQQ и SCQ су параметризовани формулама у затвореном облику за Лапласову функцију густине вероватноће, што даље омогућава брзу и једноставну процену перформанси оба квантизера [20].

3.2.2 Поређење са скаларним компандинг квантизером

Од интереса је случај када описани SQQ надмашује SCQ за исту вредност амплитуде максималног оптерећења $x_{K-1}^{\mathrm{SQQ}} = x_{K-1}^{\mathrm{SCQ}}$, те дефинишемо неопходне параметре SCQ са *N*-нивоа:

$$\left\{ \begin{array}{l} \lambda(x) = \dfrac{1}{3\sqrt{2}} \exp\left\{ -\dfrac{\sqrt{2}x}{3} \right\}; c^{\mathrm{SCQ}}(x) = \left(1 - \exp\left\{ -\dfrac{\sqrt{2}x}{3} \right\} \right), x \geq 0; \\ x_i^{\mathrm{SQQ}} = c^{\mathrm{SCQ}^{-1}}\left(\dfrac{i}{K} \right), i = 0,1,...,K-1; y_i^{\mathrm{SQQ}} = c^{\mathrm{SCQ}^{-1}}\left(\dfrac{2i-1}{2K} \right), i = 1,2,...,K; \\ x_{K-1}^{\mathrm{SQQ}} = \dfrac{3}{\sqrt{2}} \ln(K); D^{\mathrm{SCQ}} = \dfrac{9}{2N^2} - \dfrac{1 - 8\ln 8 \cdot \left(\ln\sqrt{8} - 1 \right)}{N^3}. \end{array} \right\}. \qquad (3.2.2.1)$$

У наставку показујемо да дисторзија прекорачења за SCQ није еквивалентна са дисторзијом прекорачења за SQQ. Конкретно, за амплитуду максималног оптерећења x_{K-1}^{SCQ} за SCQ, дизајниран за Лапласову функцију густине вероватноће нулте средње вредности и јединичне варијансе одређујемо укупну дисторзију и показујемо да се разликује од дисторзије SQQ дате једначином (3.2.1.21).

Укратко, изостављамо комплетан доказ и само истичемо његове главне кораке [20]:

- Позивајући се на изразе (3.2.1.15), (3.2.1.17) и (3.2.1.18), и дискусију о грануларној дисторзији SQQ, из услова $x_{K-1}^{\mathrm{SQQ}} = x_{K-1}^{\mathrm{SCQ}}$ следи $D_g^{\mathrm{SQQ}} = D_g^{\mathrm{SCQ}}$. Пажљивом анализом ових израза, можемо интуитивно одредити грануларну дисторзију као $D_{\mathrm{g}}^{\mathrm{SCQ}} = \dfrac{9}{2N^2}\left(1 - \dfrac{2}{N} \right)$.

- Поновимо срачунавање y_K^{SCQ}, где важи $c\left(y_K^{SCQ}\right)=\frac{N-1}{N}$, тако да је $y_K^{\mathrm{SCQ}}=\frac{3}{\sqrt{2}}\ln(N)$.
- Нагласимо да y_K^{SCQ} и y_K^{SQQ} имају директан утицај на дисторзију прекорачења.
- Заменом y_K^{SCQ} у (3.2.1.18) долази се до $D_{\mathrm{over}}^{\mathrm{SCQ}}$ у следећој форми:

$$D_{\mathrm{over}}^{\mathrm{SCQ}}=2\int_{x_{K-1}^{\mathrm{SQQ}}}^{x_K^{\mathrm{SQQ}}}\left(x-y_K^{\mathrm{SCQ}}\right)^2 p(x)dx=\frac{8\left[1+\ln 8\cdot\left(\ln\sqrt{8}-1\right)\right]}{N^3}. \qquad (3.2.2.2)$$

Финално, тотална дисторзија и SQNR датог SCQ се одређују као:

$$D^{\mathrm{SCQ}}=D_{\mathrm{g}}^{\mathrm{SCQ}}+D_{\mathrm{over}}^{\mathrm{SCQ}}=\frac{9}{2N^2}-\frac{1-8\ln 8\cdot\left(\ln\sqrt{8}-1\right)}{N^3},\ \mathrm{SQNR}^{\mathrm{SCQ}}=-10\log_{10}\left(D^{\mathrm{SCQ}}\right). \qquad (3.2.2.3)$$

У разматрању дисторзије прекорачења и грануларне дисторзије, прихватили смо да се велике вредности амплитуда сигнала који се компримује, јављају значајно ређе од малих. За појашњење SQQ анализе, чини се разумним прихватање монотоности суседних ћелија, која произилази из одређивања величине ћелије, као што је илустровано на сл. 3.2.1.2. Из (3.2.1.10) се може закључити да Δ_i^{SQQ} опада строго монотоно до Δ_0^{SQQ} kako *i* опада, док изворна функција густине вероватноће има коначан грануларни регион x_{K-1}^{SQQ} и $\Delta_{K-1}^{\mathrm{SQQ}}$ је ћелија коначне ширине, тј. ћелија која лежи на граници грануларног региона. Са порастом N ($N\to\infty$), а уз занемаривање дисторзије прекорачења због сужавања области прекорачења, може се приметити да се једначина (3.2.1.21) приближава Пантер-Дитеовој формули високе резолуције $D^{\mathrm{PD}}=9/(2N^2)$ [61]:

$$D_N^{\mathrm{SQQ}}{}_{|N\to\infty}=\lim_{N\to\infty}\left(\frac{9}{2N^2}\left(1-\frac{10}{9N}\right)\right)=\frac{9}{2N^2}. \qquad (3.2.2.4)$$

Дисторзија прекорачења се експлицитно може срачунати применом Бенетовог интеграла уз замену x_{K-1}^{SQQ} из (3.2.1.14):

$$D_{\mathrm{over}}^{\mathrm{B.I.}}=\frac{2\Delta^2}{12}\int_{x_{K-1}^{\mathrm{SQQ}}}^{x_K^{\mathrm{SQQ}}}\left[\frac{\partial c^{\mathrm{SQQ}}(x)}{\partial x}\right]^{-2} p(x)dx, \qquad (3.2.2.5)$$

$$D_{\text{over}}^{\text{B.I.}} = \frac{9}{N^3}. \tag{3.2.2.6}$$

Из израза (3.2.2.4) и (3.2.2.6), се лако долази до израза за грануларну дисторзију SQQ:

$$D_g^{\text{SQQ}} = D_N^{\text{SQQ}}{}_{|N\to\infty} - D_{\text{over}}^{\text{B.I.}} = \frac{9}{2N^2}\left(1-\frac{2}{N}\right). \tag{3.2.2.7}$$

који се очекивано поклапа са изразом (3.2.1.17).

Ради анализе дисторзије прекорачења, можемо упоредити формуле дате у једначинама (3.2.1.20) и (3.2.2.6), са оном у једначини (3.2.2.2) која је изведена за модел компандора. Очигледно, важи $D_{\text{over}}^{\text{SQQ}} < D_{\text{over}}^{\text{SCQ}} < D_{\text{over}}^{\text{B.I.}}$, што је заиста једна од погодних карактеристика описаног SQQ модела.

Имајући у виду да релативни компетитивни односи дефинисани као [20]:

$$\delta_{\text{over}}^{\text{SQQ/B.I.}} = \left|\frac{D_{\text{over}}^{\text{SQQ}} - D_{\text{over}}^{\text{B.I.}}}{D_{\text{over}}^{\text{SQQ}}}\right| \cdot 100 = 125[\%], \tag{3.2.2.8}$$

$$\delta_{\text{over}}^{\text{SCQ/B.I.}} = \left|\frac{D_{\text{over}}^{\text{SCQ}} - D_{\text{over}}^{\text{B.I.}}}{D_{\text{over}}^{\text{SCQ}}}\right| \cdot 100 = 3{,}93[\%]. \tag{3.2.2.9}$$

не зависе од *N*, можемо истаћи да је ово занимљиво сазнање које заслужује даљу пажњу, посебно када је значај дисторзије прекорачења израженији. Такође можемо истаћи да, све док дисторзија прекорачења није претежно занемарљива у односу на грануларну дисторзију, $D_{\text{over}}^{\text{SQQ}}$ пружа значајну предност над $D_{\text{over}}^{\text{SCQ}}$ и $D_{\text{over}}^{\text{B.I.}}$.

Преформулацијом проблема оптимизације у броју битова, као проблема предвиђања MSE дисторзије, SQNR би се могао извести директно из формула које карактеришу његове перформансе, дате у једначинама (3.2.1.21) и (3.2.1.22):

$$\text{SQNR}_{\text{Asym}}^{\text{SQQ}} = -10\log_{10}\left(D_{\text{Asym}}^{\text{SQQ}}\right), \tag{3.2.2.10}$$

$$\text{SQNR}_{\text{Exact}}^{\text{SQQ}} = -10\log_{10}\left(D_{\text{Exact}}^{\text{SQQ}}\right). \tag{3.2.2.11}$$

Да бисмо обезбедили поређење перформанси дефинише се релативна грешка:

$$\delta^{\mathrm{A}} = \left| \frac{\mathrm{SQNR}_{\mathrm{Asym}}^{\mathrm{SQQ}} - \mathrm{SQNR}_{\mathrm{Exact}}^{\mathrm{SQQ}}}{\mathrm{SQNR}_{\mathrm{Exact}}^{\mathrm{SQQ}}} \right|, \tag{3.2.2.12}$$

$$\delta^{\mathrm{PD}} = \left| \frac{\mathrm{SQNR}^{\mathrm{PD}} - \mathrm{SQNR}_{\mathrm{Exact}}^{\mathrm{SQQ}}}{\mathrm{SQNR}_{\mathrm{Exact}}^{\mathrm{SQQ}}} \right|. \tag{3.2.2.13}$$

Вредности SQNR представљене у Табели 3.2.2.1. показују да $\mathrm{SQNR}_{\mathrm{Asym}}^{\mathrm{SQQ}}$ обезбеђују значајно повећање перформанси у поређењу са $\mathrm{SQNR}^{\mathrm{PD}}$. Вреди приметити да се за $N \geq 128$, $\mathrm{SQNR}_{\mathrm{Asym}}^{\mathrm{SQQ}}$ постиже приближно исти ниво тачности као и код $\mathrm{SQNR}_{\mathrm{Exact}}^{\mathrm{SQQ}}$ или $\mathrm{SQNR}^{\mathrm{PD}}$, за исти број нивоа N, што се и интуитивно могло очекивати.

Табела 3.2.2.1. Поређење SQNR за Asym, Exact и PD приступе и преглед релетивне грешке за Asym и PD

N	$\mathrm{SQNR}_{\mathrm{Asym}}^{\mathrm{SQQ}}$[dB]	$\mathrm{SQNR}_{\mathrm{Exact}}^{\mathrm{SQQ}}$[dB]	$\mathrm{SQNR}^{\mathrm{PD}}$[dB]	δ^{A}	δ^{PD}
4	6.9224	7.1865	5.5091	0.0368	0.2334
6	9.9203	10.1329	9.0309	0.0210	0.1087
8	12.1791	12.3501	11.5297	0.0138	0.0664
10	13.9794	14.1212	13.4679	0.0100	0.0463
12	15.4734	15.5943	15.0515	0.0077	0.0348
14	16.7496	16.8547	16.3904	0.0062	0.0275
16	17.8629	17.9558	17.5503	0.0052	0.0226
18	18.8500	18.9333	18.5733	0.0044	0.0190
20	19.7367	19.8121	19.4885	0.0038	0.0163
28	22.5869	22.6416	22.4110	0.0024	0.0102
30	23.1742	23.2254	23.0103	0.0022	0.0093
32	23.7243	23.7725	23.5709	0.0020	0.0085
64	29.6675	29.6919	29.5915	0.0008	0.0034
128	35.6499	35.6622	35.6121	0.0003	0.0014
256	41.6516	41.6577	41.6327	0.0001	0.0006

Иако даје приближне вредности, дату асимптотску формулу за $\mathrm{SQNR}_{\mathrm{Asym}}^{\mathrm{SQQ}}$ одликује прилична тачност, чак и за мале вредности N, што је чини веома корисном. Друго, пошто су δ^{A} и δ^{PD} приближно мањи или једнаки 1% за $N = 10$ и $N = 30$, респективно, потврђује се коректност описаног SQQ модела за мале и средње битске брзине.

Треће, може се приметити да су вредности битских брзина R ($R = \log_2 N$) при којима релативна грешка δ^{A} нагло расте у опсегу $R \in$ [2 бит/одмерак, 3 бит/одмерак]. Напоменимо да само најмањи број нивоа, $N = 4$, показује доминантно највећу релетивну грешку $\delta^{\mathrm{A}} = 3{,}68\%$. Уз ова сазнања, лакше је пратити различита ограничења која се могу поставити битској брзини у потрази за минимизираним вредностима MSE дисторзије.

Формално, овде коришћен приступ проналажења дисторзије, директно повезан са одређивањем SQNR, ($\mathrm{SQNR} = -10\log_{10}(D)$) [2] нуди потпунији увид у утицај обе – грануларне дисторзије и дисторзије прекорачења на укупну дисторзију. Да бисмо сагледали предности модела SQQ, извршићемо анализу за неколико различитих нивоа квантизације. Релативна компетитивна предност у односу на теорију високе резолуције се може дефинисати преко перформансног добитка G^{SQQ} као:

$$G^{\mathrm{SQQ\text{-}PD}} = 10\log_{10}\left(\frac{D^{\mathrm{PD}}}{D^{\mathrm{SQQ}}}\right) = 10\log_{10}\left(\frac{9N}{9N-10}\right), \qquad (3.2.2.14)$$

док се компетитивна предност у односу на компандор, тј. SCQ, може наћи као:

$$G^{\mathrm{SQQ\text{-}SCQ}} = 10\log_{10}\left(\frac{D^{\mathrm{SCQ}}}{D^{\mathrm{SQQ}}}\right) = 10\log_{10}\left(\frac{9N - 2\left(1 - 8\ln 8\cdot\left(\ln\sqrt{8}-1\right)\right)}{9N-10}\right). \qquad (3.2.2.15)$$

Упоређујући вредности перформансног добитка из табеле 3.2.2.2. можемо доћи до сличних закључака за мале и средње брзине преноса, $R \in$ [2 бит/одмерак, 5 бит/одмерак]. У наставку упоређујемо SQNR вредности и анализирамо SQNR криве које се преклапају дуж коначног дела њихове дужине.

Имајмо на уму да је преклапање крива илустративан индикатор и стога, прецизније, наш циљ је да нагласимо снажно преклапање за $\mathrm{SQNR}_{\mathrm{Asym}}^{\mathrm{SQQ}}$ и $\mathrm{SQNR}_{\mathrm{Exact}}^{\mathrm{SQQ}}$, као и $\mathrm{SQNR}^{\mathrm{PD}}$ и $\mathrm{SQNR}^{\mathrm{SCQ}}$ за R у опсегу 2 бит/одмерак до 5 бит/одмерак (слика 3.2.2.1. (а)). Такође се са слике сл. 3.2.2.1. б) може приметити да када је R у опсегу од 3 бит/одмерак до 5 бит/одмерак, $\mathrm{SQNR}_{\mathrm{Asym}}^{\mathrm{SQQ}}$ надмашује друга асимптотска решења као $\mathrm{SQNR}^{\mathrm{PD}}$ и $\mathrm{SQNR}^{\mathrm{Hui}}$ срачунате као у [61] и [32], респективно:

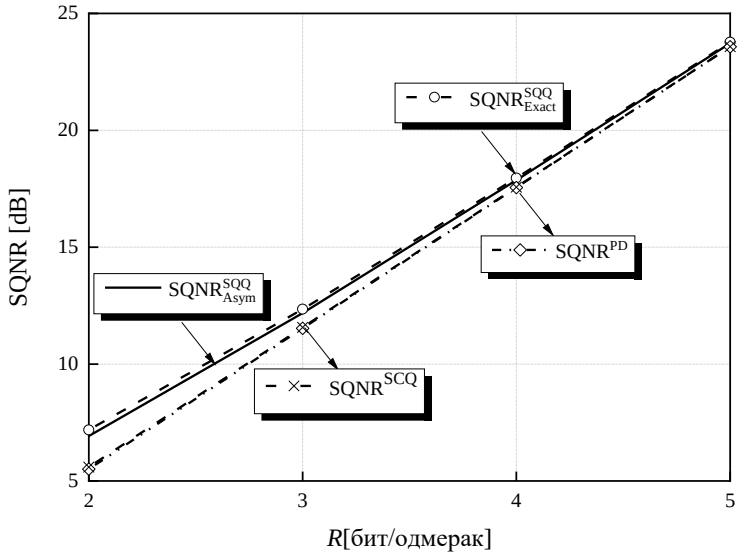
$$\mathrm{SQNR}^{\mathrm{PD}} = -10\log_{10}\left(D^{\mathrm{PD}}\right) = 10\log_{10}\left(\frac{2N^2}{9}\right), \tag{3.2.2.16}$$

$$\mathrm{SQNR}^{\mathrm{Hui}} = 10\log_{10}\left(\frac{3N^2}{2\ln^2(N)+3}\right). \tag{3.2.2.17}$$

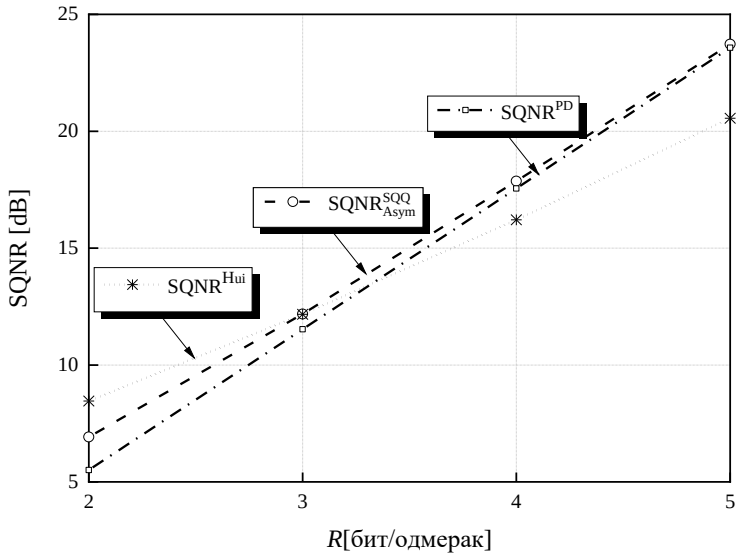
док за R = 2 бит/одмерак, $\mathrm{SQNR}^{\mathrm{Hui}}$ превазилази SQNR описаног SQQ.

Табела 3.2.2.2. SQQ перформансни добитак у односу на високо-резолуциону теорију и SCQ модел.

N	$D_{\mathrm{Asym}}^{\mathrm{SQQ}}$	D^{PD}	D^{SCQ}	$G^{\mathrm{SQQ\text{-}PD}}[\mathrm{dB}]$	$G^{\mathrm{SQQ\text{-}SCQ}}[\mathrm{dB}]$
4	0.20313	0.28125	0.27595	1.4133	1.3307
6	0.10185	0.12500	0.12342	0.8894	0.8345
8	0.06055	0.07031	0.06965	0.6494	0.6083
10	0.04000	0.04500	0.04466	0.5115	0.4787
12	0.02836	0.03125	0.03105	0.4220	0.3946
14	0.02114	0.02296	0.02284	0.3591	0.3357
16	0.01636	0.01758	0.01750	0.3125	0.2921
18	0.01303	0.01389	0.01383	0.2767	0.2585
20	0.01063	0.01125	0.01121	0.2482	0.2318
32	0.00424	0.00440	0.00438	0.1535	0.1432
64	0.00108	0.00110	0.00110	0.0760	0.0710
128	0.000272	0.000275	0.000274	0.0379	0.0353
256	0.0000684	0.0000687	0.0000686	0.0189	0.0176



а)



б)

Слика 3.2.2.1. SQNR зависност од *R* за мале и средње битске брзине: а) Примена израза (3.2.2.3), (3.2.2.10), (3.2.2.11) и (3.2.2.16). б) Примена израза (3.2.2.10), (3.2.2.11) и (3.2.2.17).

Објашњење за ово се може поткрепити чињеницом да је за тако малу брзину преноса, вредност амплитуде максималног оптерећења која је оптимизована асимптотском анализом за $N >> 1$, мање тачна за случају неуниформног квантизера, него за случају униформног, што се последично одражава на SQNR вредности ова два модела квантизера. Анализа дата у овом одељку упућује на закључак да се тотална дисторзија своди на грануларну дисторзију, све док се дисторзија прекорачења не јавља или је претежно занемарљива.

Док је у већини случајева где се квантизација традиционално користи, теорија високе резолуције ($N \to \infty$) добро оправдана, она је потпуно неоправдана за класе квантизера са малим или средњим бројевима квантизационих ћелија N. Велике вредности N доводе до прецизнијег квантовања, али резултирају мање ефикасном рачунским и меморијским ресурсима. Многи квантизери често раде на веома скромним брзинама, односно нижим или средњим битским брзинама, које су наметнуте различитим ограничења доступних ресурса. Сходно томе, верујемо да би овде дата анализа, изведена не само за веће брзине преноса ($R \geq 5$ бит/одмерак), већ и за средње битске брзине, могла бити од великог значаја [20].

3.3 Пројектовање квазилогаритамског квантизера за Лапласов извор

3.3.1 Оптимизација грануларног региона

Као што је познато, модели компресије података могу претпоставити извор информација са познатим статистичким својствима. Пратећи главни аспект кодовања и компресије података, као што је смањење броја битова за пренос и складиштење оригиналног аналогног сигнала, очигледно је да компандинг квантизери добијају велики истраживачки интерес, посебно јер су погодни са аспекта лакшег дизајна, сложености и имплементације у односу на друге класе неуниформних квантизера. Квалитет реподукције сигнала мери се укупном дисторзијом, тј. сумом грануларне и дисторзије прекорачења, које се као функције понашају супротно услед промене параметара квантизера [2]. У овом одељку описујемо квазилогаритамски квантизер прилагођен за Лапласову функцију густине вероватноће.

За квазилогаритамски квантизер Q_μ^N, дизајниран за варијансу $\sigma_\text{р}^2$, где је μ фактор компресије а N број квантизационих нивоа, компресија применом μ-закона укључује компресорску функцију $c_\mu(x)$: $\left[-x_{\max}^{\sigma_\text{р}}, x_{\max}^{\sigma_\text{р}}\right] \rightarrow \left[-x_{\max}^{\sigma_\text{р}}, x_{\max}^{\sigma_\text{р}}\right]$ [63] - [66]:

$$c_\mu(x) = \frac{x_{\max}^{\sigma_\text{р}}}{\ln(1+\mu)} \ln\left(1+\mu\frac{|x|}{x_{\max}^{\sigma_\text{р}}}\right)\text{sgn}(x), \tag{3.3.1.1}$$

$|x| \le x_{\max}^{\sigma_\text{р}}$, а $x_{\max}^{\sigma_\text{р}}$ је граница грануларног региона квазилогаритамског квантизера. За дати грануларни регион $\left[-x_{\max}^{\sigma_\text{р}}, x_{\max}^{\sigma_\text{р}}\right]$ израз за тоталну дисторзију је:

$$D\left(Q_\mu^N\right) = C\sigma_\text{р}^{\,2}\left[\frac{1}{\mu^2}\frac{x_{\max}^{\sigma_\text{р}\,2}}{\sigma_p^{\,2}} + \frac{x_{\max}^{\sigma_\text{р}}}{\sigma_\text{р}}\frac{\sqrt{2}}{\mu} + 1\right] + \sigma_p^2 \exp\left(-\frac{\sqrt{2}x_{\max}^{\sigma_\text{р}}}{\sigma_\text{р}}\right), \tag{3.3.1.2}$$

где је $C = \ln^2(\mu+1)/(3N^2)$ константа.

Анализа дата у [2] даје релацију између грануларне дисторзије, дисторзије прекорачења и укупне дисторзије. Додатно, у [2] је дата интерпретација утицаја броја

квантизационих нивоа и величине области грануларног региона на грануларну дисторзију и дисторзију прекорачења. За квантизер са N нивоа се може рећи да је глобално оптималан за дату функцију густине вероватноће уколико је укупна дисторзија, одређена за тај квантизер и претпостављену функцију густине вероватноће минимална. Приступ који укључује процену и оптимизацију грануларног региона квантизера, као и броја квантизационих нивоа, показао се као прилично користан за пројектовање скаларних квантизера за познате расподеле улазног сигнала, а који имају за циљ постизање минималне дисторзије [57], [13]. Једно хеуристичко решење за одређивање грануларног региона оптималних и асимптотски оптималних скаларних квантизера фиксне брзине је већ приказано у [57]. Такође, рад [13] се бави одређивањем горње границе грануларног региона, тј. амплитуде максималног оптерећења. У [67], [68] дата су нека решења двостепеног модела квантизације заснованог на квазилогаритамским квантизерима. Рад [69], који је укључује технику адаптације уназад, понудио је сложено, али побољшано решење за модел квазилогаритамског квантизера, који је погодан примарно за мање вредности фактора компресије.

Циљ овог одељка је процена грануларног региона уз давање израза за амплитуду максималног оптерећења $x_{\max}^{\sigma_p}$ и њено нумеричко израчунавање у случајевима где фактор компресије μ има произвољну вредност.

Ако претпоставимо да је вредност параметра μ довољно велика (на пример μ = 255 како је дефинисано стандардом G.711), добијамо следећу приближну формулу за укупна дисторзије, као у [19]:

$$D^{\mathrm{L}}\left(Q_{\mu}^{N}\right)=D\left(Q_{\mu=255}^{N}\right)=C{\sigma_{\mathrm{p}}}^{2}\left[\frac{x_{\max}^{\sigma_{\mathrm{p}}}}{\sigma_{\mathrm{p}}}\frac{\sqrt{2}}{\mu}+1\right]+\sigma_{\mathrm{p}}^{2}\exp\left(-\frac{\sqrt{2}x_{\max}^{\sigma_{\mathrm{p}}}}{\sigma_{\mathrm{p}}}\right). \qquad (3.3.1.3)$$

До минималне дисторзије се долази изједначавањем првог извода дисторзије по $x_{\max}^{\sigma_{\mathrm{p}}}$ са нулом:

$$\frac{\ln^{2}\left(1+\mu\right)}{3N^{2}}\frac{1}{\mu}-\exp\left\{-\frac{\sqrt{2}x_{\max}^{\sigma_{\mathrm{p}},\mathrm{L}}}{\sigma_{p}}\right\}=0, \qquad (3.3.1.4)$$

те одређивањем вредности амплитуде максималног оптерећења из:

$$x_{\max}^{\sigma_p,\mathrm{L}} = \frac{\sigma_\mathrm{p}}{\sqrt{2}} \ln\left(\frac{3N^2\mu}{\ln^2(1+\mu)} \right). \tag{3.3.1.5}$$

За одређивање $x_{\max}^{\sigma_p}$, а из услова минимума дисторзије дате у (3.3.1.3), потребна је примена итеративног метода [19]:

$$x_{\max}^{\sigma_p\,(i)} = \frac{\sigma_p}{\sqrt{2}} \ln\left(\frac{3N^2\mu}{\ln^2(1+\mu)} \frac{1}{1+\frac{\sqrt{2}}{\mu}\frac{x_{\max}^{\sigma_p\,(i-1)}}{\sigma_p}} \right) = \frac{\sigma_p}{\sqrt{2}} \ln\left(\frac{\mu}{C} \frac{1}{1+\frac{\sqrt{2}}{\mu}\frac{x_{\max}^{\sigma_p\,(i-1)}}{\sigma_p}} \right). \tag{3.3.1.6}$$

Како би се убрзала процена $x_{\max}^{\sigma_p}$, за покретање итеративног поступка се користи $x_{\max}^{\sigma_\mathrm{p},\mathrm{L}}$ дат у (3.3.1.5):

Корак 1. Иницијализација $x_{\max}^{\sigma_\mathrm{p}(0)} = x_{\max}^{\sigma_\mathrm{p},\mathrm{L}}$

$$x_{\max}^{\sigma_\mathrm{p}(1)}\Big|_{x_{\max}^{\sigma_\mathrm{p}(0)} = x_{\max}^{\sigma_\mathrm{p},\mathrm{L}}} = \frac{\sigma_\mathrm{p}}{\sqrt{2}} \ln\left(\frac{\frac{\mu}{C}}{1+\frac{1}{\mu}\ln\left(\frac{\mu}{C}\right)} \right). \tag{3.3.1.7}$$

Корак 2. Срачунавање $x_{\max}^{\sigma_\mathrm{p}(2)}$ коришћењем (3.3.1.6) и прекид итеративног метода. Процена релативне грешке је:

$$\delta[\%] = \left| \frac{x_{\max}^{\sigma_p\,(2)} - x_{\max}^{\sigma_p\,(1)}}{x_{\max}^{\sigma_p\,(2)}} \right| \times 100\,, \tag{3.3.1.8}$$

уз проверaву да ли је релативна грешка мања од $5\cdot10^{-3}$.

У складу са (3.3.1.5) и (3.3.1.7), за $x_{\max}^{\sigma_\mathrm{p}} = k\sigma_\mathrm{p}$ долазимо до израза за мултипликатор *k*:

$$x_{\max}^{\sigma_\mathrm{p}} = k\sigma_\mathrm{p}\,, \tag{3.3.1.9}$$

$$k=\begin{cases}\dfrac{1}{\sqrt{2}}\ln\left(\dfrac{3N^2\mu}{\ln^2(1+\mu)}\right), \text{за } x_{\max}^{\sigma_p,\mathrm{L}} \\ \dfrac{1}{\sqrt{2}}\ln\left(\dfrac{\dfrac{3N^2\mu}{\ln^2(1+\mu)}}{1+\dfrac{1}{\mu}\ln\left(\dfrac{3N^2\mu}{\ln^2(1+\mu)}\right)}\right), \text{за } x_{\max}^{\sigma_p(1)}\end{cases}. \qquad (3.3.1.10)$$

Како би сагледали употребу квазилогаритамског квантизера, пројектованог за варијансу σ_p^2, у широком динамичком опсегу варијансе сигнала на његовом улазу σ_q^2, уводи се коефицијент скалирања ρ, који повезује произвољну варијансу сигнала који треба квантовати σ_q^2 и варијансу за коју је квантизер пројектован σ_p^2:

$$\rho=\sigma_q/\sigma_p. \qquad (3.3.1.11)$$

Дисторзија овог квантизера се одређује из:

$$D\left(Q_\mu^N\right)=\sigma_q^2\frac{\ln^2(1+\mu)}{3N^2}\left(\frac{k^2}{\mu^2\rho^2}+\frac{\sqrt{2}k}{\mu\rho}+1\right), \qquad (3.3.1.12)$$

а SQNR из:

$$\mathrm{SQNR}=10\log_{10}\left(\frac{\sigma_q^2}{D\left(Q_\mu^N\right)}\right)=-10\log_{10}\left(C\left(\frac{a^2}{\mu^2}+\frac{\sqrt{2}a}{\mu}+1\right)+\exp\left\{-\sqrt{2}a\right\}\right), \qquad (3.3.1.13)$$

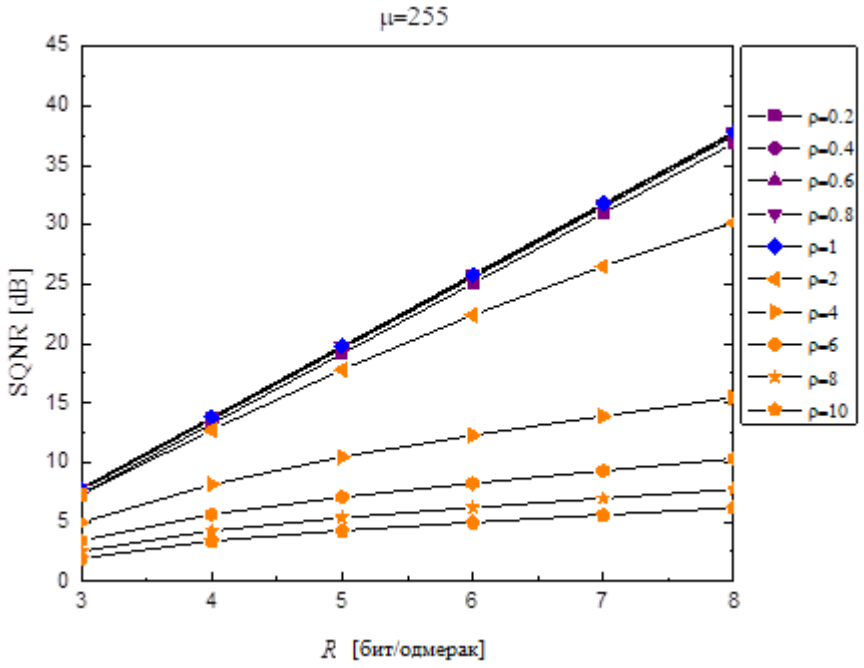
где је $a = k / \rho$.

Наиме, показује се да се избором параметра μ и R, а у складу са захтевима минималног SQNR, могу балансирати перформансе квазилогаритамског квантизера, истовремено постижући додатну оптимизацију битске брзине.Описани метод процене грануларног региона одликује једноставна иницијализација и брза конвергенција обзиром да се креће од формуле у затвореном облику за амплитуду максималног оптерећења квазилогаритамског квантизера пројектованог за Лапласов извор произвољне варијансе.

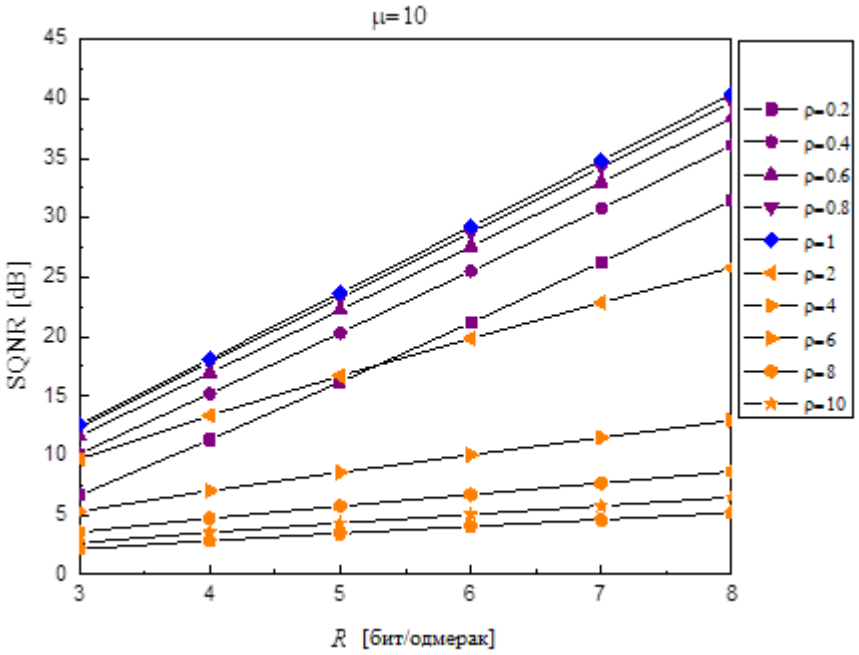
Табела 3.3.1.1. Приказ амплитуде максималног оптерећења и релативне грешке њеног одређивања.

Фактор компресије μ	R [бит/одмерак]	$x_{\max}^{\sigma_p(2)}$	$x_{\max}^{\sigma_p(1)}$	$x_{\max}^{\sigma_p(0)}$	$\delta[\%]$
μ =255	4	6.1699	6.1698	6.1937	0.0015
	6	8.1230	8.1229	8.1542	0.0014
	8	10.0763	10.0761	10.1147	0.0014
μ =100	4	5.7363	5.7357	5.7914	0.0089
	6	7.6790	7.6783	7.7519	0.0086
	8	9.6222	9.6214	9.7124	0.0083
μ =10	4	4.7285	4.7058	5.0892	0.4790
	6	6.5856	6.5606	7.0497	0.3785
	8	8.4552	8.4292	9.0102	0.3077

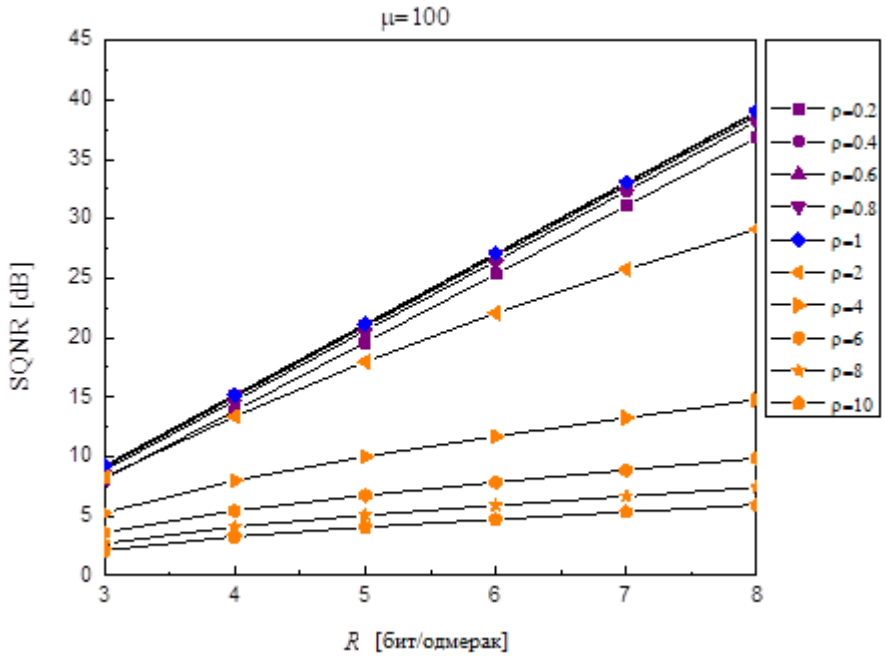
Како би се одредила процена након само једне итерације, за различите вредности коефицијента скалирања μ = 255, μ = 100, μ = 10 и број квантизационих нивоа N (N = 16, N = 64 и N = 256), одређује се релативна грешка (у складу са (3.3.1.8)) чији је преглед дат у табели 3.3.1.1. а која је до 0.5 %, у најнеповољнијем случају. Зависност SQNR од μ је дата за различите специфициране вредности ρ као и за μ = 255, μ = 10 и μ = 100 (сл.3.3.1.1.- сл.3.3.1.3.).Примећујемо да за μ = 10 и R = 5 бит/одмерак или 6 бит/одмерак, SQNR се креће у опсегу нешто већем од 15 dB када се коефицијенти скалирања креће од ρ = 0.2 до ρ =2. За μ = 100 и R =3 бит/одмерак или R = 4 бит/одмерак, SQNR се креће у опсегу већем од 10 dB за исти распон коефицијента скалирања ρ = 0.2 до ρ = 2. При вишим битским брзинама, нпр. за $R \geq 8$ бит/одмерак, уз промену коефицијента скалирања ρ = 0.2 до ρ = 2, постиже се већи SQNR од 20 или 30 dB, за μ = 10 или μ = 100, респективно.



Слика 3.3.1.1. SQNR зависност од коефицијента скалирања ρ за μ = 255 и R = 3 - 8 [бит/одмерак].



Слика 3.3.1.2. SQNR зависност од коефицијента скалирања ρ за μ = 10 и R = 3 - 8 [бит/одмерак].



Слика 3.3.1.3. SQNR зависност од коефицијента скалирања ρ за μ = 100 и R = 3 - 8 [бит/одмерак].

3.3.2 Итеративно одређивање границе грануларног региона

За јединичну варијансу, амплитуда максималног оптерећења се може наћи као:

$$x_{\max} = \frac{1}{\sqrt{2}} \ln\left[\frac{1}{C\left(\frac{\sqrt{2} x_{\max}}{\mu^2} + \frac{1}{\mu} \right)} \right], \tag{3.3.2.1}$$

што захтева итеративно одређивање [20]:

$$x_{\max}^{(i)} = \frac{1}{\sqrt{2}} \ln\left(\frac{\mu}{C\left(1 + \frac{\sqrt{2} x_{\max}^{(i-1)}}{\mu} \right)} \right). \tag{3.3.2.2}$$

Вреди напоменути да су у [70] изведене три аналитичке процене оптималне амплитуде максималног оптерећења. Будући да је задатак оптималне процене амплитуде максималног оптерећења испуњен минимизирањем MSE дисторзије, једна од процена је истакнута као посебно погодна за практичну примену [70]. Да бисмо поједноставили процену $x_{\max}$, иницијализујемо итеративни алгоритам користећи амплитуду максималног оптерећења $t_{\max}^{(3)}$ из [70] и примењујемо следеће кораке:

Корак 1. Иницијализација $t_{\max}^{(3)} = \frac{3}{\sqrt{2}} \ln\left[N + 1 \right]$

$$x_{\max}^{(1)}\Big|_{x_{\max}^{(0)} = t_{\max}^{(3)}} = \frac{1}{\sqrt{2}} \ln\left(\frac{\frac{\mu^2}{C}}{\mu + 3\ln\left(N + 1 \right)} \right) = \frac{1}{\sqrt{2}} \ln\left(\frac{\mu}{C} \right) + \frac{1}{\sqrt{2}} \ln\left(\frac{\mu}{\mu + 3\ln\left(N + 1 \right)} \right). \tag{3.3.2.3}$$

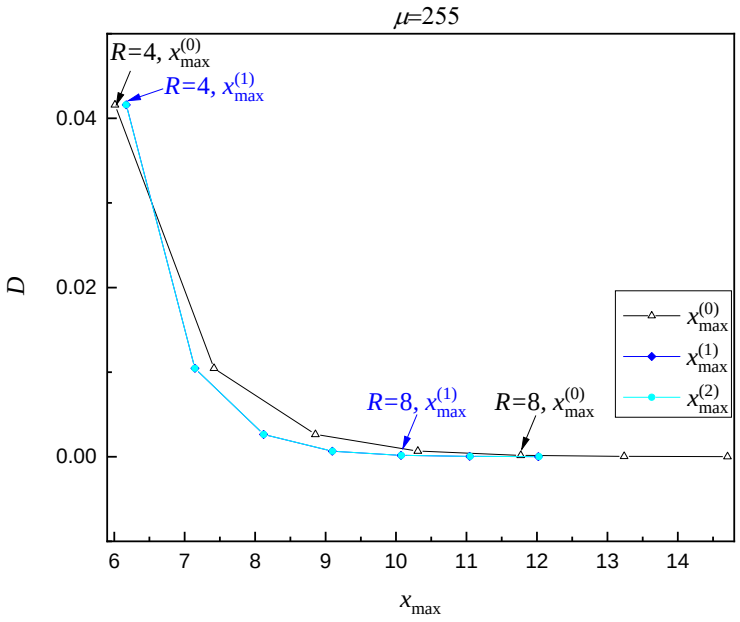
Корак 2. Срачунавање $x_{\max}^{(2)}$:

$$x_{\max}^{(2)} = \frac{1}{\sqrt{2}} \ln \left(\frac{\frac{\mu^2}{C}}{\mu + \ln \left(\frac{\frac{\mu^2}{C}}{\mu + 3\ln(N+1)} \right)} \right). \qquad (3.3.3.4)$$

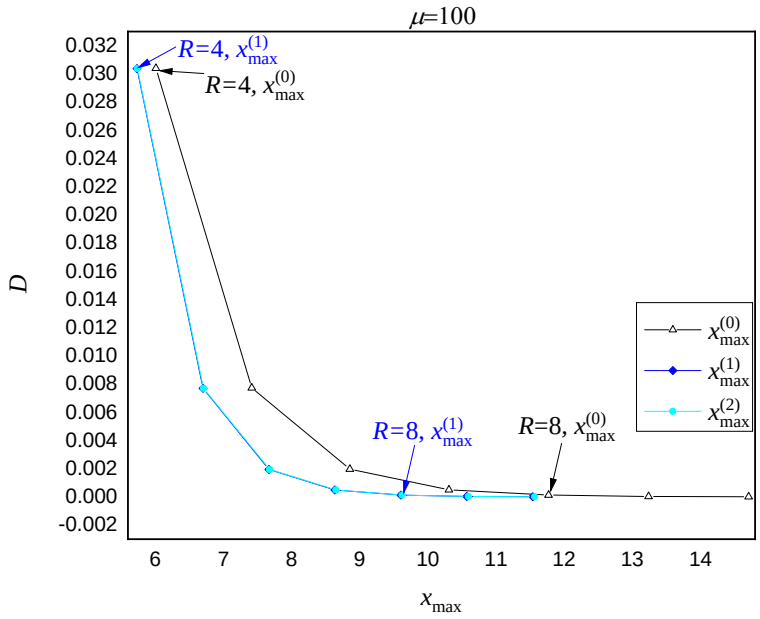
Корак 3. Процена релативне грешке:

$$\delta[\%] = \left| \frac{x_{\max}^{(2)} - x_{\max}^{(1)}}{x_{\max}^{(2)}} \right| \times 100 . \qquad (3.3.2.5)$$

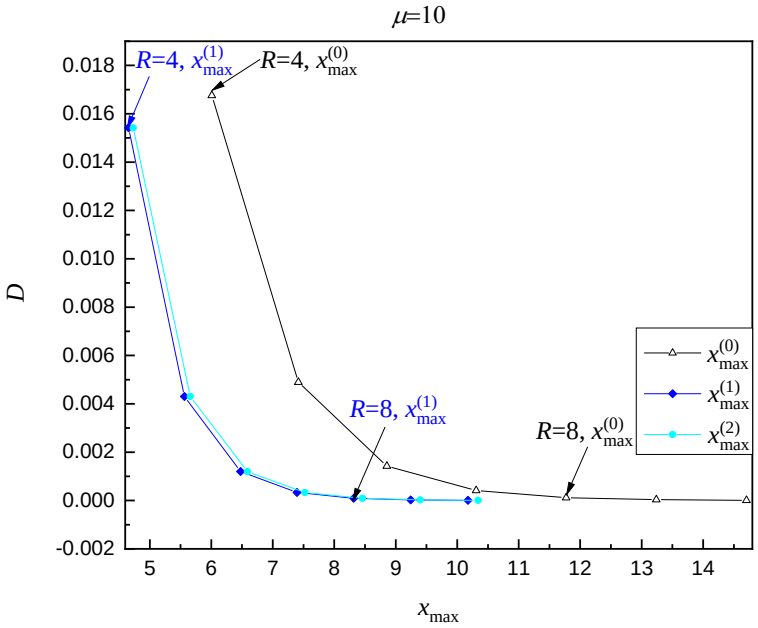
На сл. 3.3.2.1. - сл.3.3.2.3. приказана је дисторзија за карактеристичне вредности фактора компресије μ (μ = 255, μ = 100, μ = 10), у функцији од $x_{\max}^{(0)}$, $x_{\max}^{(1)}$ и $x_{\max}^{(2)}$. Маркиране вредности на криви су вредности $x_{\max}^{(0)}$, $x_{\max}^{(1)}$ и $x_{\max}^{(2)}$ срачунате су за различите бројеве квантизационих нивоа N (N = 16, 32, 64, 128, 256, 512, 1024).



Слика 3.3.2.1. Зависност дисторзије од $x_{\max}^{(0)}$, $x_{\max}^{(1)}$ и $x_{\max}^{(2)}$ за μ = 255 и R = 4 - 10 [бит/одмерак].



Слика 3.3.2.2. Зависност дисторзије од $x_{max}^{(0)}$, $x_{max}^{(1)}$ и $x_{max}^{(2)}$ за μ = 100 и R = 4 - 10 [бит/одмерак].



Слика 3.3.2.3. Зависност дисторзије од $x_{max}^{(0)}$, $x_{max}^{(1)}$ и $x_{max}^{(2)}$ за μ = 10 и R = 4 - 10 [бит/одмерак].

У наставку је дата анализа преклапања крива и изводимо закључке из пресека и делимичног преклапања дуж коначног дела њихове дужине. Ваља нагласити да је преклапање криве илустративан индикатор јер је конвергенција итеративног процеса боља када је преклапање веће. Коришћене боје крива помажу у разликовању кривих укључених у пресек, док је мешање боја довољно да се уочи скоро потпуно преклапање за $x_{\max}^{(2)}$ и $x_{\max}^{(1)}$ крива за $\mu = 255$.

Слика 3.3.2.1. за $\mu = 255$ приказује криве за $x_{\max}^{(2)}$ и $x_{\max}^{(1)}$ које секу $x_{\max}^{(0)}$ криву за R између 4 - 5 бит/одмерак. Даље, може се приметити добро преклапање $x_{\max}^{(2)}$ и $x_{\max}^{(1)}$ кривих за $\mu = 100$. Такође је приметно да се за $\mu = 10$ постиже делимично преклапање за $x_{\max}^{(2)}$ и $x_{\max}^{(1)}$ дуж коначног дела њихове дужине за R у распону од 7 - 10 бит/одмерак. Наиме, ако је почетна вредност $\overset{\equiv}{x}_{\max}^{(0)}$ за описани итеративни метод једнака $x_{\max}^{a} = 1/\sqrt{2} \cdot \ln(\mu / C)$, која се добија из приближне формуле у затвореном облику за амплитуду максималног оптерећења квазилогаритамског квантизера датог у [20], вредност нивоа $\overset{\equiv}{x}_{\max}^{(1)}$ се може срачунати као:

$$\overset{\equiv}{x}_{\max}^{(1)}\Big|_{\overset{\equiv}{x}_{\max}^{(0)} = x_{\max}^{a}} = \frac{1}{\sqrt{2}} \ln\left(\frac{\frac{\mu}{C}}{1 + \frac{1}{\mu}\ln\left(\frac{\mu}{C}\right)} \right) = \frac{1}{\sqrt{2}} \ln\left(\frac{\mu}{C}\right) + \frac{1}{\sqrt{2}} \ln\left(\frac{\mu}{\mu + \ln\left(\frac{\mu}{C}\right)} \right), \qquad (3.3.2.6)$$

$$\overset{\equiv}{x}_{\max}^{(2)} = \frac{1}{\sqrt{2}} \ln\left(\frac{\frac{\mu}{C}}{1 + \frac{1}{\mu}\ln\left(\frac{\frac{\mu}{C}}{1 + \frac{1}{\mu}\ln\left(\frac{\mu}{C}\right)} \right)} \right). \qquad (3.3.2.7)$$

У табелама 3.3.2.1. – 3.3.2.3. је дат приказ итеративно одређених амлитуда максималног оптрећења за предложене иницијализације $t_{\max}^{(3)}$ и $x_{\max}^{a}$ а за $\mu = 255$, $\mu = 10$ и $\mu = 100$.

Табела 3.3.2.1. Итеративно одређивање x_{max} за две различите иницијализације за μ = 255.

μ=255	$x_{max}^{(0)}$	$x_{max}^{(1)}$	$x_{max}^{(2)}$	$\equiv x_{max}^{(0)}$	$\equiv x_{max}^{(1)}$
R = 4	6.0102	6.1705	6.1699	6.19369	6.1698
R = 5	7.4172	7.1454	7.1465	7.17395	7.1464
R = 6	8.8552	8.1203	8.1231	8.15420	8.1229
R = 7	10.3092	9.0952	9.0997	9.13446	9.0995
R = 8	11.7714	10.0700	10.0763	10.11472	10.0761
R = 9	13.2376	11.0449	11.0529	11.09498	11.0528
R = 10	14.7059	12.0198	12.0296	12.07524	12.0294

Табела 3.3.2.2. Итеративно одређивање x_{max} за две различите иницијализације за μ = 100.

μ=100	$x_{max}^{(0)}$	$x_{max}^{(1)}$	$x_{max}^{(2)}$	$\equiv x_{max}^{(0)}$	$\equiv x_{max}^{(1)}$
R = 4	6.0102	5.7337	5.7363	5.7914	5.7357
R = 5	7.4172	6.7011	6.7076	6.7717	6.7070
R = 6	8.8552	7.6685	7.6791	7.7519	7.6784
R = 7	10.3092	8.6359	8.6507	8.7322	8.6498
R = 8	11.7714	9.6036	9.6224	9.7124	9.6214
R = 9	13.2376	10.5714	10.5942	10.6927	10.5931
R = 10	14.7059	11.5394	11.5660	11.6730	11.5649

Табела 3.3.2.3. Итеративно одређивање x_{max} за две различите иницијализације за μ = 10.

μ=10	$x_{max}^{(0)}$	$x_{max}^{(1)}$	$x_{max}^{(2)}$	$\equiv x_{max}^{(0)}$	$\equiv x_{max}^{(1)}$
R = 4	6.0102	4.6542	4.7316	5.0892	4.7058
R = 5	7.4172	5.5622	5.6591	6.0694	5.6313
R = 6	8.8552	6.4756	6.5900	7.0497	6.5606
R = 7	10.3092	7.3940	7.5239	8.0299	7.4934
R = 8	11.7714	8.3172	8.4604	9.0102	8.4292
R = 9	13.2376	9.2445	9.3993	9.9905	9.3676
R = 10	14.7059	10.1754	10.3403	10.9707	10.3084

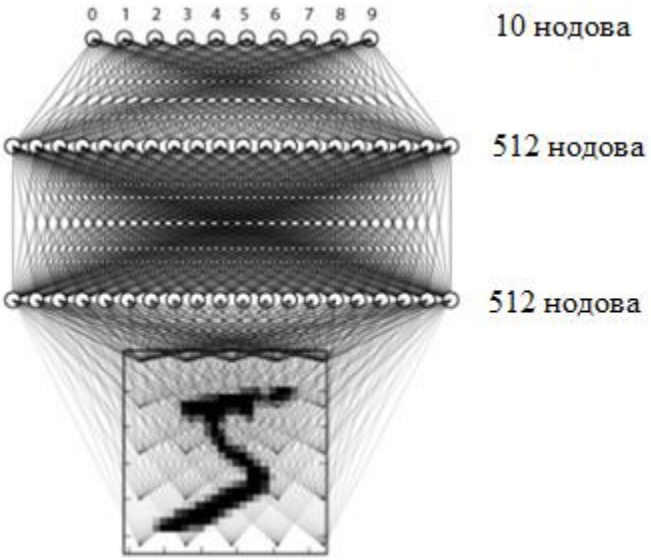
4. Пројектовање нискобитних униформних квантизера и њихова примена код неуронских мрежа

4.1 Модел неуронске мреже за примену квантизера у пост-тренинг квантизацији

За експерименталну процену, тј. евалуацију перформанси нискобитног униформног квантизера примењеног у квантизацији параметара (тежина) неуронске мреже у фази након тренинга у истраживањима спроведеним у радовима [21] - [24], као тестни модел неуронске мреже користили смо трослојни потпуно повезани FC (*Fully Connected*) NN модел (укратко, наш тестни NN модел). Блок дијаграм приказан на слици 4.1.1.1. приказује наш експериментални процес евалуације.

Слика 4.1.1.1. Блок дијаграм за експерименталну процену перформанси квантизације после тренинга.

Прва одабрана архитектура NN модела је MLP, и прилично је једноставна, а састоји се од три потпуно повезана слоја (видети слику 4.1.1.2.), јер је циљ анализа утицаја дизајна квантизера на тачност MLP модела, али и постизање највеће могуће тачности на самом скупу података. Штавише, увек ћемо претпостављати исту архитектуру за све даље описане моделе квантизера, како би обезбедили коректно поређење резултата у овом и наредном поглављу.

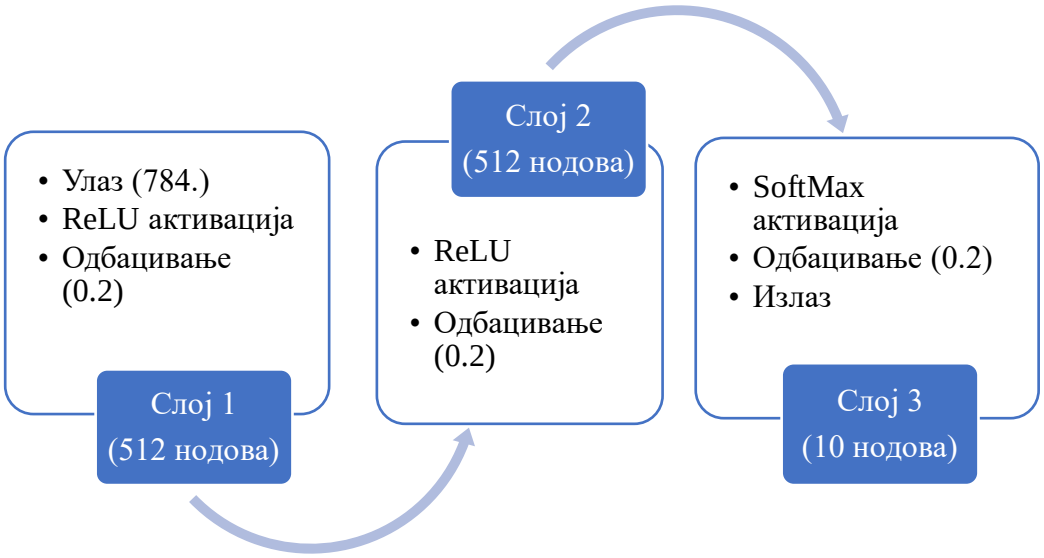


Слика 4.1.1.2. MLP модел за препознавање слика руком писаних цифара.

Друга одабрана архитектура NN модела је CNN, која садржи један додатни конволуциони слој. Оба NN модела су обучена на MNIST скупу података за препознавање слика које садрже једну руком писану цифару, који се састоји од 60 000 црно-белих слика руком писаних цифара од 0 до 9 [71]. Конкретно, MNIST скуп података се састоји од 70 000 црно-белих слика, са вредностима интензитета пиксела у опсегу [0-255]. Скуп података је подељен на 60 000 слика за обуку и 10 000 слика за тестирање, док све слике имају једнаке димензије 28x28 пиксела [71]. Слике за MLP су предефинисане у једнодимензионалне векторе од 784 (28x28,) елемента да би одговарале облику који је прихвата први слој NN, док се за правилан CNN улаз додаје једна додатна димензија која представља канал. Последњи корак претходне припреме улаза је његова нормализација у опсег [0-1], дељењем сваког интензитета пиксела слике са највећом могућом вредношћу за амплитуду црно-беле слике, тј. са 255.

MLP модел се састоји од два скривена и једног излазног слоја, са укупно 3 потпуно повезана (густа) слоја (види слику 4.1.1.2.). Скривени слојеви се састоје од 512 нодова, док се излазни слој састоји од 10 нодова, где сваки излазни нод процењује вероватноћу

да је MLP излаз било која написана цифра (од 0 до 9). Оба скривена слоја користе ReLU активациону функцију и регуларизацију одбацивања (*Dropout*) која поставља 20% излаза скривених слојева на 0 [72]. На излазном слоју се користи *SoftMax* активациона функција, која дефинише вероватноће да улаз припада било којој од 10 могућих класа (видети слику 4.1.1.3.)



Слика 4.1.1.3. Трослојни FC NN модел за препознавање слика руком писаних цифара.

CNN модел се састоји од једног конволуционог слоја, праћеног ReLU активацијом, слојем сажимања максималном вредношћу (*max pooling*) и слојем поравнања (*flatten layer*), чији се излаз доводи у претходно описани MLP са 2 скривена FC слоја и излазним слојем. Конволуциони слој садржи 16 филтара са величином језгра 3 × 3, док слој сажимања користи прозор величине 2 × 2. Излаз слоја сажимања је даље трансформисан у једнодимензионални вектор који се доводи на FC густи слој. Једина разлика између претходно описаног MLP-а и густих слојева који се користе у CNN је у проценту одбацивања, који је у случају CNN постављен на 50 %, како би се додатно спречило прекомерно тренирање (*overfitting*) FC слојева. Дакле, CNN модел се састоји од конволуционог слоја, иза кога следе два скривена слоја и један излазни слој са укупно 1 652 906 параметара (тежина) који се могу обучити. Обука се изводи за MNIST

скуп података, на исти начин као и код MLP, у укупно 10 епоха, док је број слика које се обрађују одједном (*batch size*) 128.

За оба модела NN, обука, анализа тачности и квантизација, су имплементирани у програмском језику Python [73]. Након обуке NN модела, тако одређене тежине се чувају као 32-битне презентације са покретним зарезом, што представља пуну прецизност у TensorFlow оквиру. MLP модел постиже тачност од 98.1 % на сету за валидацију, које је постигнута након 10 епоха обуке. CNN модел постиже већу тачност од 98.89 % на сету за валидацију, постигнуту такође након 10 епоха обуке. Први NN модел (MLP) састоји се од 669 706 параметара, а други NN модел (CNN) од 1 652 906 који се доводе на улаз предложеног нискобитног униформног квантизера. По одређивању квантованих репрезентација тежина NN модела, исте се могу вратити у QNN модел. Циљ наших истраживања у [21] - [24] је процена перформанси QNN за претпостављени модел нискобитног квантизера.

Резултати су такође дати за оба наведена модела, MLP и CNN, који су тренирани на Fashion-MNIST скупу података [74]. Fashion-MNIST је скуп података који се састоји од 28×28 црно-белих слика 70 000 модних производа из 10 категорија, са 7 000 слика по категорији [74]. Као и код MNIST сета података, сет за обуку има 60 000 слика, док сет за теститање има 10 000 слика. У [75] је истакнуто да иако скуп података Fashion-MNIST представља изазовнији задатак за класификацију у поређењу са MNIST скупом података, али употреба MNIST скупа података и даље не застарева, будући да MNIST скуп података пружа могућност брзе провере различитих алгоритама и модела квантизације код неуронских мрежа. За оба NN модела, обука, процена тачности и квантизација су имплементирани у програмском језику Python [73], [76].

- MLP модел постиже тачност од 98.1 % на MNIST сету за валидацију и 88.96 % на Fashion-MNIST сету за валидацију, а након 10 епоха обуке.
- CNN модел постиже већу тачност од 98.89 % на MNIST сету за валидацију и 91.53 % на Fashion-MNIST сету за валидацију, такође након 10 епоха обуке.

Након тренинга разматраног NN модела врши се нормализација тежина, које се нормализују тако да имају нулту средњу вредност и јединичну стандардну девијацију. Након квантовања нормализованох тежина врши се денормализација, тако да се квантоване тежине враћају у првобитан опсег вредности. Другим речима, денормализација се врши након процеса квантизације и пре учитавања квантованих тежина назад у NN модел (детаљније погледати одељак 4.2.5). Након што се квантоване тежине поново учитају у модел, процењујемо перформансе и нискобитног униформног квантизера изражене SQNR вредношћу и тачношћу QNN. Слично процењујемо перформансе квантизера који се прилагођавају амплитудској динамици тежина по слојевима NN.

У наредним одељцима овог поглавља дати су:

- Анализа утицаја избора кључног параметра дизајна двобитног и тробитног униформног квантизера на SQNR и резултујућу тачност QNN за MNIST скуп података за моделе скаларних квантизере дате у [21] - [24].
- Одговор на питање да ли се избор кључног параметра, тј. амплитуде максималног оптерећења нискобитног униформног квантизера, одражава и на SQNR и на тачност QNN у истој мери.
- Одговор на питање да ли је могуће применити најједноставнију униформну квантизацију за квантовање обучених тежина NN модела са битском брзином R=2 бит/одмерак и R=3 бит/одмерак, уз очување тачности NN модела у великој мери.
- Резултате који показују да се даље побољшање тачности може постићи потпуно новим приступом заснованим на по слојевима адаптираној униформној квантизацији тежина за дати NN модел, са екпериментални резултатима за R=2 бит/одмерак.

4.2 Перформансе нискобитне униформне и по слојевима адаптиране униформне квантизације у пост-тренинг фази из перспективе избора грануларног региона

4.2.1 Дизајн симетричног двобитног униформног квантизера за Лапласов извор и његова верзија адаптирана по слојевима

Поновимо, квантизација специфицира пресликавање вредности на улазу квантизера у дискретни скуп од N нивоа квантизације. За меру грешке настале применом квантизације, обично се узима средње-квадратна грешка дисторзије [2].

$$D = E\left[\left(X - Q_N\left(X\right)\right)^2\right]. \tag{4.2.1.1}$$

Циљ квантизације је да се минимизира грешка између квантованих вредности ($Q_N(X)$) и оригиналних вредности на улазу у квантизер (X), за дато N, тј. за дати број битова потребних за представљање ових врдности, тј. података $R = \log_2 N$ [2]. Циљ компресије је смањење битске брзине, пошто коришћење ниже битске брзине смањује меморијске и рачунске захтеве, али и повећава грешку услед квантизације [2]. Због ових сукобљених захтева, квантизација је веома интригантна област истраживања. Конкретно, као што је већ поменуто, избор самог модела квантизера и његовог грануларног региона (кључни параметар сваког квантизера), утичу на укупну грешку услед квантизације, те ћемо најпре специфицирати кључни параметар симетричног квантизера Q_N са N нивоа.

Опсег амплитуде улазног сигнала се дели на грануларни регион $\Re_g$ и регион прекорачења $\Re_o$ (погледати сл. 4.2.1.1. за симетрични двобитни униформни квантизер). У случају симетричног квантизера, што је случај са квантизером који је предмет анализе овог одељка, ова два региона су раздвојена симетричним амплитудама максималног оптерећења означеним са $-x_{\max}$ и $x_{\max}$, респективно. Грануларни регион $\Re_g$ је дефинисан са:

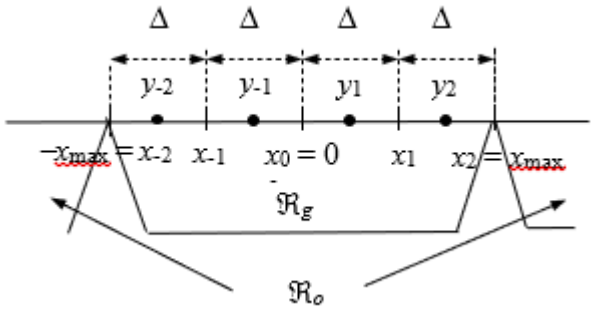
$$\Re_g = \bigcup_{i=-N/2}^{-1} \Re_i \cup \bigcup_{i=1}^{N/2} \Re_i = [-x_{\max}, x_{\max}], \tag{4.2.1.2}$$

и састоји се од N непреклапајућих квантизационих ћелија ограничених ширина, где је i-та ћелија дефинисана као:

$$\Re_i = \{x \mid x \in [-x_{\max}, x_{\max}], Q_N(x) = y_i\}, \Re_i \cap \Re_j = \varnothing,\ i \neq j\,. \tag{4.2.1.3}$$

$\{\Re_i\}_{i=-N/2}^{-1}$ и $\{\Re_i\}_{i=1}^{N/2}$ означавају грануларне ћелије у негативним и позитивним амплитудним регионима, који су симетрични. Код симетричних квантизера главни скуп параметара квантизера може бити преполовљен, јер само позитивне или апсолутне вредности параметара квантизера треба одредити и сачувати. Ова симетрија важи и за квантизационе ћелије у региону прекорачења, односно за пар ћелија неограничене ширине у региону прекорачења $\Re_o$, који је дефинисан као [21]:

$$\Re_0 = \{x \mid x \notin [-x_{\max}, x_{\max}], Q_N(x) = y_{N/2}, x > 0 \vee Q_N(x) = y_{-N/2}, x < 0\}\,. \tag{4.2.1.4}$$



Слика 4.2.1.1. Грануларни регион $\Re_g$ и регион прекорачења $\Re_o$ симетричног двобитног UQ.

У нискобитној квантизацији, веома мали број битова по одмерку се користи за представљање података који се квантују (мање или једнако 3 бит/одмерак) [2]. Ако су ћелије квантизера једнаке ширине, онда је квантизер униформан, што је случај са овде разматраним квантизером. Кодна књига двобитног униформног квантизера, $Y \equiv \{y_{-2}, y_{-1}, y_1, y_2\} \subset \mathbb{R}$, садржи $N = 4$ репрезентациона нивоа, означена са y_i (видети сл. 4.2.1.1.) који су позиционирани на средини квантизационих ћелија:

$$y_i = \frac{\left(x_{i-1} + x_i\right)}{2},\ y_{-i} = -y_i,\ i \in \left\{1, 2\right\}. \tag{4.2.1.5}$$

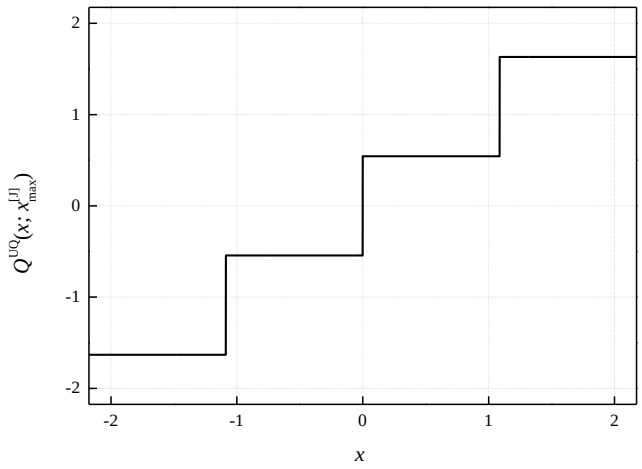
Границе ћелије су еквидинстантне и одређене са:

$$x_i = i\Delta,\ x_{-i} = -x_i,\ i \in \left\{0, 1, 2\right\}. \tag{4.2.1.6}$$

Δ је величина корака, тј. ширина ћелија двобитног униформног квантизера дата са:

$$\Delta = \frac{2x_{\max}}{N} = \frac{x_{\max}}{2}. \tag{4.2.1.7}$$

Подсетимо се да $x_{\max}$ означава амплитуду максималног оптерећења, а уједно је и кључни параметар посматраног двобитног униформног квантизера. Из израза (4.2.1.5) - (4.2.1.7) може се закључити да $x_{\max}$ у потпуности одређује границе ћелија x_i и репрезентационе нивое y_i двобитног униформног квантизера. Стога уводимо следећу нотацију преносне карактеристике симетричних двобитног униформног квантизера, Q^{UQ} (x; $x_{\max}$) (погледати сл. 4.2.1.2. где је преносна карактеристика симетричног двобитног униформног квантизера представљена за $x_{\max} = x_{\max}^{[J]} = 2.1748$, где ознака [J] потиче од имена аутора [2]), као што је дато у [21].



Слика 4.2.1.2. Преносна карактеристика симетричног двобитног UQ, $Q^{UQ}(x; x_{\max}^{[J]})$.

Због симетрије неограничене Лапласове функције густине вероватноће нулте средње вредности и варијансе $\sigma^2 = 1$ за коју описујемо дизајн униформног квантизера, границе ћелија и репрезентациони нивои су симетрични у односу на средњу вредност.

$$p(x) = \frac{1}{\sqrt{2}\sigma} \exp\left\{ -\frac{\sqrt{2}x}{\sigma} \right\}. \tag{4.2.1.8}$$

Да бисмо одредили укупну дисторзију нашег симетричног двобитног униформног квантизера, $D^{\mathrm{UQ}}=D_{\mathrm{g}}{}^{\mathrm{UQ}}+D_{\mathrm{o}}{}^{\mathrm{UQ}}$, почињемо са основном дефиницијом дисторзије, датом изразом (1) [2], где су грануларна дисторзија $D_{\mathrm{g}}{}^{\mathrm{UQ}}$ и дисторзија прекорачења $D_{\mathrm{o}}{}^{\mathrm{UQ}}$ симетричног двобитног униформног квантизера дефинисане као:

$$D_{\mathrm{g}}^{\mathrm{UQ}} = 2\sum_{i=1}^{2} \int_{x_{i-1}}^{x_i} (x - y_i)^2 \, p(x) dx, \tag{4.2.1.9}$$

$$D_{\mathrm{o}}^{\mathrm{UQ}} = 2\int_{x_2}^{\infty} (x - y_2)^2 \, p(x) dx. \tag{4.2.1.10}$$

Уводимо даље да је $x_3 = \infty$, где x_3 означава горњу границу интеграла у изразу (4.2.1.10) па се укупна дисторзија нашег симетричног двобитног униформног квантизера може одредити као:

$$D^{\mathrm{UQ}} = 2\sum_{i=1}^{3} \int_{x_{i-1}}^{x_i} x^2 p(x) dx - 4\left(\sum_{i=1}^{2} y_i \int_{x_{i-1}}^{x_i} x p(x) dx + y_2 \int_{x_2}^{\infty} x p(x) dx \right) + 2\left(\sum_{i=1}^{2} y_i{}^2 \int_{x_{i-1}}^{x_i} p(x) dx + y_2{}^2 \int_{x_2}^{\infty} p(x) dx \right) \tag{4.2.1.11}$$

$$D^{\mathrm{UQ}} = \mathrm{I}^{\mathrm{I}} - 4\left(\sum_{i=1}^{2} y_i \mathrm{I}_i^{\mathrm{II}} + y_2 \mathrm{I}_3^{\mathrm{II}} \right) + 2\left(\sum_{i=1}^{2} y_i{}^2 \mathrm{I}_i^{\mathrm{III}} + y_2{}^2 \mathrm{I}_3^{\mathrm{III}} \right), \tag{4.2.1.12}$$

где за претпостављену функцију густине вероватноће дату изразом (4.2.1.8), изводимо:

$$\mathrm{I}^{\mathrm{I}} = 2\sum_{i=1}^{3} \int_{x_{i-1}}^{x_i} x^2 p(x) dx = \sigma^2, \tag{4.2.1.13}$$

$$\mathrm{I}_i^{\mathrm{II}} = \int_{x_{i-1}}^{x_i} xp(x)dx = \frac{1}{2}\left[\left(x_{i-1} + \frac{\sigma}{\sqrt{2}}\right)\exp\left\{-\frac{\sqrt{2}x_{i-1}}{\sigma}\right\} - \left(x_i + \frac{\sigma}{\sqrt{2}}\right)\exp\left\{-\frac{\sqrt{2}x_i}{\sigma}\right\}\right], i = 1,2,3, \quad (4.2.1.14)$$

$$\mathrm{I}_i^{\mathrm{III}} = \int_{x_{i-1}}^{x_i} p(x)dx = \frac{1}{2}\left[\exp\left\{-\frac{\sqrt{2}x_{i-1}}{\sigma}\right\} - \exp\left\{-\frac{\sqrt{2}x_i}{\sigma}\right\}\right], i = 1,2,3. \quad (4.2.1.15)$$

Замена израза (4.2.1.13) и (4.2.1.15) у (4.2.1.12) даје:

$$D^{\mathrm{UQ}} = \sigma^2 + {y_1}^2 - \sqrt{2}\sigma y_1 + \left[\left(y_2 - y_1\right)\left(y_2 + y_1 - 2x_1 - \sqrt{2}\sigma\right)\right]\exp\left\{-\frac{\sqrt{2}x_1}{\sigma}\right\}. \quad (4.2.1.16)$$

Узимајући у обзир једначине (4.2.1.5) - (4.2.1.7) долазимо до следећег израза за укупну дисторзију нашег симетричног двобитног униформног квантизера када је улаз Лапласова функција густине вероватноће нулте средње вредности и произвољне варијансе σ^2:

$$D^{\mathrm{UQ}} = \sigma^2 + \frac{{x_{\max}}^2}{16} - \frac{\sqrt{2}}{4}\sigma x_{\max}\left(1 + 2\exp\left\{-\frac{\sqrt{2}\,x_{\max}}{2\,\sigma}\right\}\right). \quad (4.2.1.17)$$

Као и у бројним радовима из области квантизације (на пример, [12] - [14], [30], [39], [57], [58], [62],) анализу спроводимо за случај прилагођене варијансе, где се варијанса улазног сигнала и варијанса за коју је квантизер дизајниран подударају. Стога, даље претпостављамо јединичну варијансу, $\sigma^2 = 1$, тако да долазимо до следеће формуле за дисторзију нашег симетричног двобитног униформног квантизера за улаз са Лапласовом функцијом густине вероватноће нулте средње вредности и јединичне варијансе [21]:

$$D^{\mathrm{UQ}}\Big|_{\sigma^2=1} = 1 + \frac{{x_{\max}}^2}{16} - \frac{\sqrt{2}}{4}x_{\max}\left(1 + 2\exp\left\{-\frac{\sqrt{2}\,x_{\max}}{2}\right\}\right). \quad (4.2.1.18)$$

Затим дефинишемо теоријски SQNR као:

$$\mathrm{SQNR}_{\mathrm{th}}^{\mathrm{UQ}} = 10\log_{10}\left(\frac{\sigma^2}{D^{\mathrm{UQ}}}\right), \quad (4.2.1.19)$$

који срачунавамо за $\sigma^2 = 1$. Теоријски добијене вредности поредићемо са експериментално оствареним SQNR, одређеним за стварне тежине обучене NN, дефинисане као:

$$\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{UQ}} = 10\log_{10}\left(\frac{\frac{1}{W}\sum_{j=1}^{W} w_j^2}{\frac{1}{W}\sum_{j=1}^{W}\left(w_j - w_j^{\mathrm{UQ}}\right)^2}\right), \tag{4.2.1.20}$$

где w_j, $j = 1,2,...,W$ означавају тежине пре квантизације, представљене у FP32 формату, w_j^{UQ}, $j = 1,2,...,W$ су тежине квантоване описаним симетричним двобитним униформним квантизером, а W је укупан број тежина. Да бисмо направили разлику између теоријских и експерименталних резултата користимо x за сигнал/податке који се квантују и узимају за прорачун теоријске вредности SQNR, док користимо w за податке (тежине NN модела) који се квантују и узимају за прорачун експерименталне вредности SQNR. Коначно, можемо дефинисати правило квантизације за описане симетричне двобитне униформне квантизације тежина. За описани симетрични двобитни униформни квантизер можемо дефинисати следећу преносну карактеристику:

$$Q^{\mathrm{UQ}}\left(x; x_{\max}\right) = \begin{cases} \mathrm{sgn}(x)\left(\left\lfloor |x|/\Delta \right\rfloor + 1/2\right)\Delta, & |x| \le x_{\max} \\ \mathrm{sgn}(x)\left(x_{\max} - \Delta/2\right), & |x| > x_{\max} \end{cases}, \tag{4.2.1.21}$$

а узимајући у обзир израз (4.2.1.7), можемо дефинисати правило квантизације за описану симетричну двобитну униформну квантизацију тежина као:

$$w_j^{\mathrm{UQ}} = \begin{cases} \mathrm{sgn}(w_j)\left(\left\lfloor 2|w_j|/x_{\max} \right\rfloor + 1/2\right)x_{\max}/2, & |w_j| \le x_{\max} \\ \mathrm{sgn}(w_j)\left(x_{\max} - x_{\max}/4\right), & |w_j| > x_{\max} \end{cases}. \tag{4.2.1.22}$$

Посматрајући изразе (4.2.2.18) - (4.2.2.22), може се предвидети да је постављање одговарајуће вредности амплитуде максималног оптерећења кључно за постизање најбољег могућег учинка примењеног модела квантизера. Чак и у овом једноставном случају двобитне униформне квантизације, $x_{\max}$ се не може аналитички одредити тако

да обезбеди минималну дисторзију. Тачније, $x_{\max}$ који је одредио *Jayant* [2] је резултат нумеричке оптимизације дисторзије, док је *Hui* аналитички дошао до следећег израза за $x_{\max}$ симетричног асимптотски оптималног униформног квантизера са *N*-нивоа, дизајнираног за велике битске брзине и улазни сигнал са Лапласовом функцијом густине вероватноће, нулте средње вредности и јединичне варијанса [62]:

$$x_{\max}^{[\mathrm{H}]} = \sqrt{2}\ln(N). \tag{4.2.1.23}$$

У наставку одређујемо преносне карактеристике и правило квантизације за други модел квантизера који ћемо описати, и то LWUQ, такође дат у раду [21], а који је састављен од *M* униформних квантизера са амплитудама максималног оптерећења прилагођеним динамици амплитуде тежина на сваком од *M* слојева:

$$Q^{\mathrm{UQ}_{L_i}}\left(x; x_{\max}^{L_i}\right) = \begin{cases} \mathrm{sgn}(x)\left(\left\lfloor 2|x| / x_{\max}^{L_i} \right\rfloor + 1/2\right) x_{\max}^{L_i} / 2, & |x| \le x_{\max}^{L_i} \\ \mathrm{sgn}(x)\left(x_{\max}^{L_i} - x_{\max}^{L_i} / 4\right), & |x| > x_{\max}^{L_i} \end{cases}, L_i = 1, 2, ..., M\ , \tag{4.2.1.24}$$

$$w_j^{\mathrm{UQ}_{L_i}} = \begin{cases} \mathrm{sgn}(w_j)\left(\left\lfloor 2|w_j| / x_{\max}^{L_i} \right\rfloor + 1/2\right) x_{\max}^{L_i} / 2, & |w_j| \le x_{\max}^{L_i} \\ \mathrm{sgn}(w_j)\left(x_{\max}^{L_i} - x_{\max}^{L_i} / 4\right), & |w_j| > x_{\max}^{L_i} \end{cases}, L_i = 1, 2, ..., M\ . \tag{4.2.1.25}$$

LWUQ је предложен како би утврдили да ли додатно адаптирање амплитуде максималног оптерећења по слојевима NN модела може обезбедити побољшање у погледу тачности NN и SQNR. Разлика између LWUQ и униформног квантизера лежи само у слојевитој адаптацији амплитуде максималног оптерећења, тако да у последњим двема једначинама имамо $x_{\max}^{L_i}$, тј. амплитуде максималног оптерећења униформног квантизера које су прилагођене појединачној динамици амплитуда тежина на сваком од *M* слојева (L_i, i = 1, 2, …, *M*) описаног NN модела. Сходно томе, можемо специфицирати експериментални SQNR за сваки од *M* слојева као:

$$\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{UQ}_{L_i}} = 10\log_{10}\left(\frac{\frac{1}{W_{L_i}}\sum_{j=1}^{W_{L_i}} w_j^2}{\frac{1}{W_{L_i}}\sum_{j=1}^{W_{L_i}}\left(w_j - w_j^{\mathrm{UQ}_{L_i}}\right)^2}\right), L_i = 1, 2, ..., M\ , \tag{4.2.1.26}$$

и за све слојеве као:

$$\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{LWUQ}} = 10\log_{10}\left(\frac{\frac{1}{M}\sum_{L_i=1}^{M}\frac{1}{W_{L_i}}\sum_{j=1}^{W_{L_i}} w_j^2}{\frac{1}{M}\sum_{L_i=1}^{M}\frac{1}{W_{L_i}}\sum_{j=1}^{W_{L_i}}\left(w_j - w_j^{\mathrm{UQ}_{L_i}}\right)^2}\right). \qquad (4.2.1.27)$$

где W_{L_i} означава број тежина на слоју L_i.

4.2.2 Преглед нумеричких и експерименталних перформансних индикатора за двобитни униформни квантизер и LWUQ

У овом одељку анализирамо тачност представљеног NN модела након униформне квантизације тежина (QNN модел), где је униформни квантизер пројектован за битску брзину од R = 2 бит/одмерак [21]. Као што је већ поменуто, NN модел са FP32 форматом тежина, постиже тачност од 98.1 % на MNIST скупу података за валидацију. Како компресија тежина неизбежно деградира тачност нашег модела QNN, да бисмо смањили степен деградације, пожељно је у потпуности искористити доступну битску брзину одржавајући квантоване тежине што је могуће ближе оригиналним.

Као што је поменуто раније у овом поглављу, перформансе квантизера се процењују одређивањем SQNR у dB (децибелима), док се перформансе модела QNN описују његовом тачношћу постигнутом на сету за валидацију. Како у литератури није дата директна веза између тачности QNN и SQNR, пожељно је ово испитати што је инспирисало наша истраживања дата у раду [21]. Интуитивна очекивања су да избор ширине грануларне области, $\Re_g$, има велики утицај на тачност модела. Како би потврдили интуицију, анализирали смо вишеструке изборе ширине $\Re_g$, односно дизајнирали наше двобитне униформне квантизере за различите $\Re_g$. Међу многим посматраним случајевима одабрали смо да представимо 4 репрезентативна случаја дизајна двобитног униформног квантизера, са нумеричким резултатима датим у табели 4.2.2.1. Уочени случајеви се разликују по избору ширине $\Re_g$ имплементираног квантизера, а у табели 4.2.2.1. су дати перформансни параметри оба модела, за QNN модел и наш двобитни униформни квантизер. Резултате слојевите адаптације $\Re_g$, где се двобитни LWUQ примењује у компресији тежина NN, а $\Re_g$ је одређен према статистици тежина сваког од слојева NN, представљени су у табели 4.2.2.2. Примена LWUQ се разматра само за случајеве 1 и 2, пошто у ова два случаја одређујемо $\Re_g$ квантизера према статистици тежина. Напоменимо да се дизајн описаног квантизера у случајевима 3 и 4 може сматрати зависним само од битске брзине.

У прва два случаја, двобитни униформни квантизер је дизајниран према статистици оригиналних нормализованих тежина модела. Посматрајући статистику тежина трениране NN, утврдили смо минималне и максималне вредности тежина у оригиналном FP32 формату, које износе $x_{min} = -7.063787$ и $x_{max} = 4.8371024$. За потребе поређења, случајеви 3 и 4 имплементирају максималне вредности x_{max} из литературе [62] и [2] одређене за $R = 2$ бит/одмерак, које не зависе од статистике улазних тежина. Заједно са експериментално постигнутим вредностима SQNR, израчунатим коришћењем израза (4.2.1.20), табела 4.2.2.1. такође представља теоријске SQNR вредности, срачунате из једначина (4.2.1.18) и (4.2.1.19) за описани двобитни униформни квантизер, дизајниран за различите ширине $\Re_g$ и Лапласову функцију густине вероватноће нулте средње вредности и јединичне варијансе. Први случај представља симетрични двобитни униформни квантизер дизајниран за максималну амплитуду нормализованих тежина нашег тренираног NN модела. У овом случају је $\Re_g$ дефинисан као $[-x_{max}, x_{max}]$ тј. одговара интервалу $[-4.8371024, 4.8371024]$. Описани дизајн двобитног униформног квантизера за случај 1 користи шири $\Re_g$ у поређењу са случајевима 3 и 4, обухватајући 99.988 % вредности тежина. Наиме, у првом случају постоји само 0.012 % тежина изван $\Re_g$, а које се налазе унутар интервала $[-7.063787, 4.8371024]$.

Табела 4.2.2.1. SQNR и тачност модела код различитих дизајна двобитних униформних квантизера.

$x_{min} = -7.063787$, $x_{max} = 4.8371024$, $x_{max}^{[H]} = 1.9605$, $x_{max}^{[J]} = 2.1748$	**Случај 1**	**Случај 2**	**Случај 3**	**Случај 4**
	$\Re_g$	$\Re_g$	$\Re_g$	$\Re_g$
	$[-x_{max}, x_{max}]$	$[x_{min}, -x_{min}]$	$[-x_{max}^{[H]}, x_{max}^{[H]}]$	$[-x_{max}^{[J]}, x_{max}^{[J]}]$
$SQNR_{ex}^{UQ}$[dB]	2.8821	-1.2402	**8.7676**	8.7639
$SQNR_{th}^{UQ}$[dB]	1.9360	-2.0066	6.9787	**7.0707**
Тачност [%]	**96.97**	94.58	96.34	96.74
Припадност $\Re_g$ [%]	99.988	100	94.787	96.691

Међу посматраним случајевима, са квантизером имплементираним у случају 1 постигнута је највећа тачност NN модела током валидације, и то 96.97 %, што указује да помоћу квантизације деградирамо тачност нашег NN модела за само 1.13 %. Иако наш предлог у случају 1 постиже највећу тачност, вредност SQNR од 2.8821 dB је прилично мала у поређењу са случајевима 3 и 4.

Случај 2 је најнеповољнији у погледу тачности модела и SQNR, а због прешироког $\Re_g$ избора двобитног униформног квантизера, дизајнираног да покрије интервал $[-x_{\min}, x_{\min}]$ (за наш конкретни NN модел [−7.063787, 7.063787]). У поређењу са случајем 1, $\Re_g$ у случају 2 укључује 100% вредности нормализованих тежина али и узак интервал [4.8371024, 7.063787], у који не пада ниједна врсдност тежина, што га чини непотребно широким. Поред тога, пошто тежине модела показују тенденцију да прате Лапласову функцију фустине вероватноће (видети слику 4.2.2.1.), већина тежина је симетрично концентрисана око нуле, што чини средину нашег грануларног региона најважнијим делом за квантизацију. Прилагођавањем ширине $\Re_g$ пуном опсегу вредности тежина за задато R деградирамо тачност квантизације у делу $\Re_g$ у коме се налази већина вредности тежина, што доводи до веома велике дисторзије коју уноси описани униформни квантизер, док SQNR има чак негативну вредност од −1.2402 dB, а теоријски SQNR износи око −2 dB. Пошто је тачност нашег модела QNN у случају 2 најнижа (94.58 %), долазимо до закључка да усвајање $\Re_g$ описаног квантизера тако да обухвата све вредности нормализованих тежина, тј. да покрива интервал $[-x_{\min}, x_{\min}]$, као што се претпостављало нпр. [15] и [40], заправо и није погодно решење за квантизацију тежина после тренинга, тачније, за нашу циљану анализу.

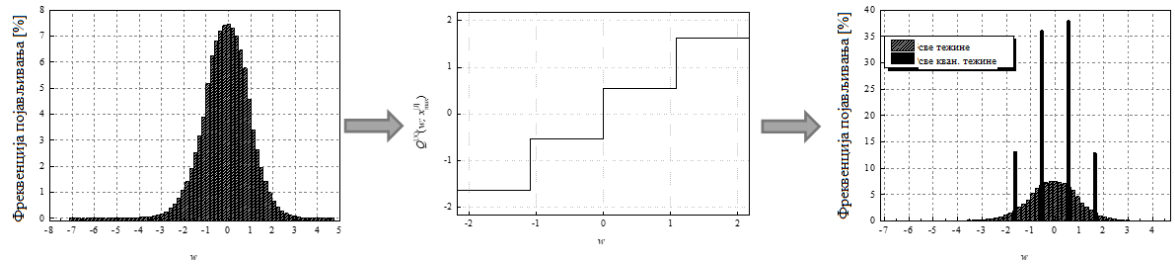
Случајеви 3 и 4 користе вредности амплитуде максималног оптерећења из литературе, које су дефинисали *Hui* [62] и *Jayant* [2]. Случај 3 користи аналитички изведен израз (4.2.1.23) за амплитуду максималног оптерећења симетричног асимптотски оптималног униформног квантизера са N нивоа, дизајнираног за улазни сигнал са Лапласовом функцијом густине вероватноће нулте средње вредности и јединичне варијансе. Како користимо двобитни квантизер, број репрезентационих нивоа ($N = 2^R$) износи $N = 4$, а амплитуда максималног оптерећења дефинисна изразом

(4.2.1.23) је $x_{\max}^{[H]} = 1.9605$. Може се приметити да је $\Re_g$ [−1.9605, 1.9605] знатно ужи у односу на претходно дефинисане $\Re_g$ и да обухвата 94.787 % свих вредност инормализованих тежина. Ово позитивно утиче на вредност SQNR која за наведени случај достиже свој максимум и износи 8.7676 dB. Тачност QNN у овом случају је такође очувана са 96.34 %, што резултира деградацијом од 0.63 %, у поређењу са случајем 1, као и 1.76 % у поређењу са оригиналним моделом NN. Коначно, случај 3 такође потврђује да избор $x_{\max}$ за који се постиже максималан SQNR нужно не подразумева и максималну тачност модела за тај исти $x_{\max}$, јер је постигнута највећа експериментална вредност SQNR (највеће вредности су подебљане), док је тачност QNN нижа него у случајевима 1 и 4. Случај 4 има сличне перформансне карактеристике као и претходни случај 3, уз имплементацију нешто ширег $\Re_g$, са максималном амплитудом оптерећења $x_{\max}^{[J]} = 2.1748$ [2]. Ова вредност је одређена нумеричком оптимизацијом за двобитни униформни квантизер, пројектован за Лапласову функцију густине вероватноће са нултом средњом вредношћу и јединичном варијансом. У конкретном случају $\Re_g$ обухвата 96.691 % тежина, а експериментални SQNR износи 8.7639 dB. Тачност QNN је нешто боља у односу на случај 3 и износи 96.74 %, док су SQNR вредности скоро идентичне. Тачност је и даље нижа од оне остварене у случају 1 и то за 0.23 %. У случају 3, занимљиво је приметити да је експериментални SQNR већи од теоријски оствареног и то за 1.7889 dB. Као и у случају 3, теоријски одређена вредност SQNR у случају 4 је нижа од експерименталне приближно за 1.8 dB, због сличног дизајна квантизера у ова два случаја. Наиме, у експерименталној анализи се квантују нормализоване тежине из ограниченог скупа могућих вредности , у нашем случају из интервала [−7.063787, 4.8371024], док се у теоријској анализи претпоставља квантизација вредности из бесконачног скупа вредности за Лапласов извор, што резултира повећањем дисторзије, односно смањењем теоријске вредности SQNR. Као потврда овог закључа може се приметити да шири $\Re_g$, тј. случај 2, показује најмању девијацију теоријских и експерименталних резултата. Један од начина смањења одступања постиже се скалирањем $\Re_g$ одређеном константом, као што је уређено у [77].

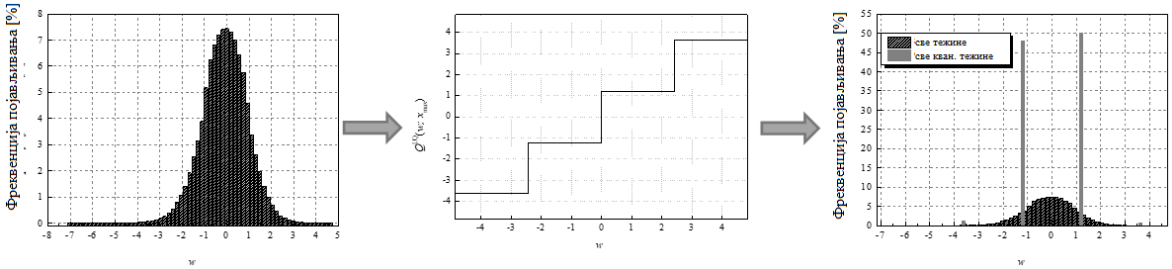
Табела 4.2.2.2. SQNR и тачност модела код различитих двобитних LWUQ дизајна.

	Случај 1	**Случај 2**
	$\mathfrak{R}_g$	$\mathfrak{R}_g$
$(x_{max}^{L1}, x_{max}^{L2}, x_{max}^{L3}) = (4.5150, 4.8371, 3.6784)$	$[-x_{max}^{L1}, x_{max}^{L1}]$	$[x_{min}^{L1}, -x_{min}^{L1}]$
$(x_{min}^{L1}, x_{min}^{L2}, x_{min}^{L3}) = (-7.0638, -5.4354, -6.1979)$	$[-x_{max}^{L2}, x_{max}^{L2}]$	$[x_{min}^{L2}, -x_{min}^{L2}]$
	$[-x_{max}^{L3}, x_{max}^{L3}]$	$[x_{min}^{L3}, -x_{min}^{L3}]$
$SQNR_{ex}^{UQ\ L1}$[dB]	3.1340	-1.7588
$SQNR_{ex}^{UQ\ L2}$[dB]	3.4507	2.2826
$SQNR_{ex}^{UQ\ L3}$[dB]	8.3642	4.6137
$SQNR_{ex}^{LWUQ}$[dB]	**3.3145**	-0.374
Тачност [%]	**97.26**	93.55

У циљу даљег побољшања перформанси и квантизера и модела QNN, разматрамо LWUQ, чије су перформансе представљене у табели 4.2.2.2. Адаптацијом $\mathfrak{R}_g$ ширине квантизера за сваки слој, повећавамо остварену SQNR вредност, што је делом и очекивано. Адаптација по слојевима иде у прилог перформансном побољшању модела QNN само у случају 1, где је тачност модела повећана за 0.29 %, уз значајно повећан SQNR. Случај 2 је већ представљен као пример за неповољан избор $\mathfrak{R}_g$ квантизера, тако да је слојевита адаптација чак умањила перформансе модела (тачност, али не и SQNR) у овом случају. Ово се може сматрати обликом пропагације грешке, пошто примењујемо грубу квантизацију у најзначајнијој области расподеле тежина, тј. у првом слоју, што је даље наглашено адаптацијом по слојевима. Перформансе квантизера се такође могу проценити посматрањем нормализованог хистограма квантованих тежина, који се изводи за униформни квантизер и LWUQ.



а)



б)

Слика 4.2.2.1. Нормализовани хистограм свих нормализованих тежина (FP32);

Преносна карактеристика симетричног двобитног UQ;

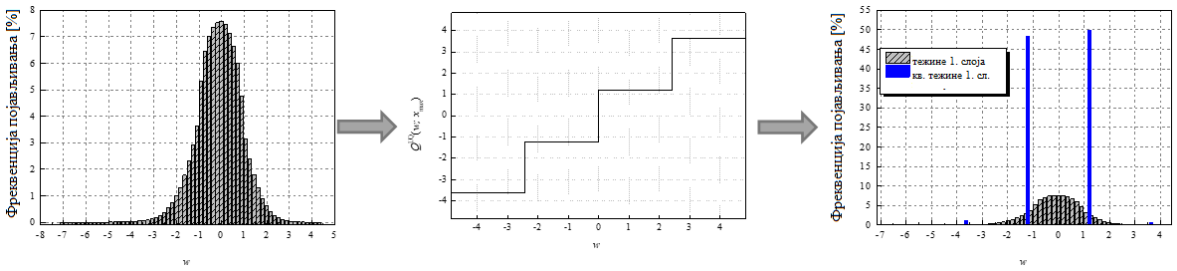
спојени нормализовани хистограм FP32 тежина и униформно квантован

а) Случај 4 (xmax[J] = 2.1748) б) Случај 1 (xmax = 4.8371024).

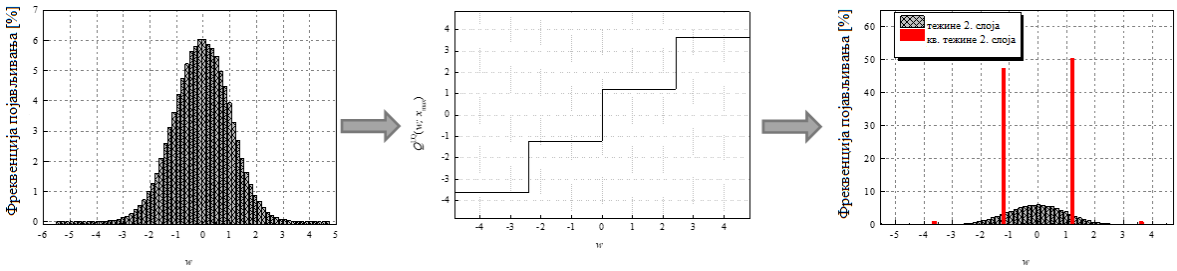
Слика 4.2.2.1. представља нормализоване хистограме свих тежина у FP32 формату, преносне карактеристике симетричног двобитног униформног квантизера и спојене нормализоване хистограме FP32 тежина и униформно квантоване за случајеве 1 и 4.

Може се приметити да у случају 4, равномерније користимо све репрезентационе нивое, док у случају 1 углавном користимо само 2 репрезентациона нивоа $\pm y_1$. Из хистограма смо утврдили за случај 1 да је 98.0657 % свих тежина представљено са $\pm y_1$, док у случају 4 тај проценат износи 74.0703 %. Дакле, у случају 1, статистички најважнији део улазних података, који се налази око нуле, тј. око средње вредности, грубо је квантован коришћењем само 2 репрезентациона нивоа, док се у случају 4 то ради коришћењем сва 4 доступна репрезентациона нивоа који су распоређени у ужем

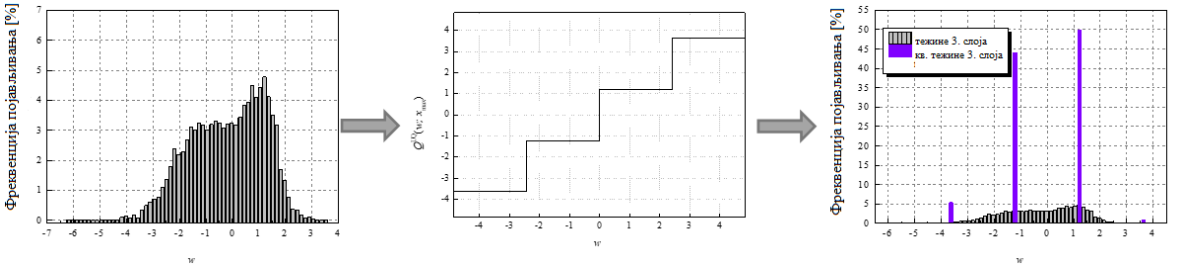
$\Re_g$. Коначно, приказани хистограми потврђују и појашњавају зашто у случају 4 добијамо много веће вредности SQNR од 8.7639 dB у поређењу са случајем 1, где остварени SQNR износи само 2.8821 dB. Занимљиво је да случај 1 пружа највећу тачност QNN, потврђујући да се тачност не ослања на највећу SQNR вредност, већ на што погоднији избор $\Re_g$ квантизера, као у случају 1.



а)



б)

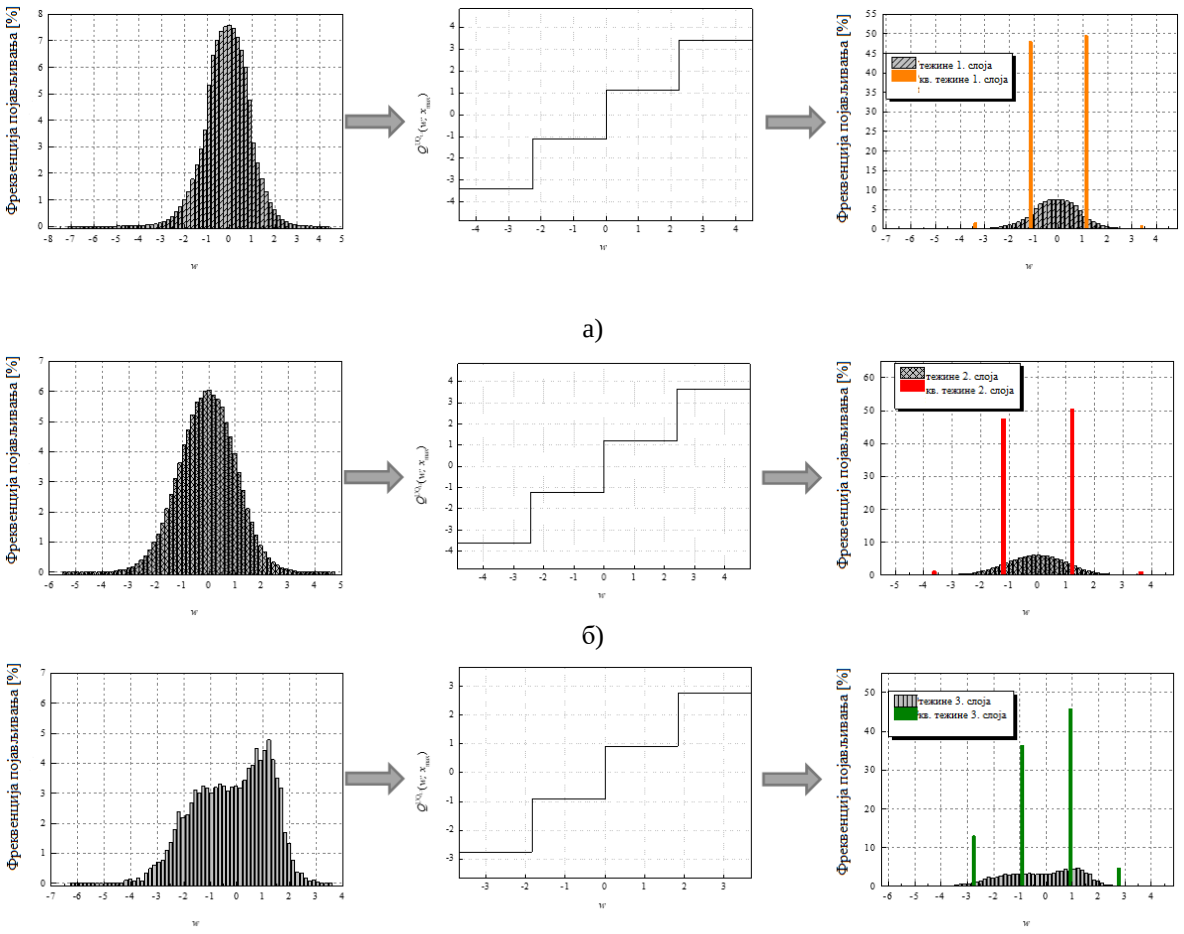


ц)

Слика 4.2.2.2. Нормализован хистограм нормализованих тежина (FP32) из а) слоја 1, б) слоја 2, ц) слоја 3;

Преносна карактеристика симетричног двобитног UQ за случај 1;

Нормализовани хистограм FP32 и униформно квантованих тежина из а) слоја 1, б) слоја 2, ц) слоја 3.



а)

б)

ц)

Слика 4.2.2.3. Нормализован хистограм нормализованих тежина (FP32) из а) слоја 1, б) слоја 2, ц) слоја 3;

По-слоју-прилагођена преносна карактеристика симетричног двобитног UQ за случај 1 и а) слој 1, б) слој 2, ц) слој 3;

FP32 по слојевима и равномерно квантоване тежине из а) слоја 1, б) слоја 2, ц) слоја 3.

Слика 4.2.2.2. представља слојевиту анализу тежина NN модела након квантизације, као и преносне карактеристике двобитног униформног квантизера примењеног на сваки слој описаног NN модела. Преносне карактеристике квантизера су идентичне, пошто двобитни униформни квантизер није адатиран појединачним слојевима. Посматрајући хистограме крајње лево, може се приметити да тежине модела NN пре квантизације имају различите расподеле у различитим слојевима. Ово значи да би

адаптација квантизера према расподели појединачних слојева заправо повећала SQNR након квантизације.

Адаптација квантизације по појединачним слојевима модела NN је представљена на слици 4.2.2.3. Може се приметити да у овом случају, уз различите расподеле тежина по слојевима, имамо различите преносне карактеристике униформних квантизера примењених на различите слојеве QNN. Тако за представљени избор $\Re_g$ случај 1, још увек имамо велику експлоатацију два репрезентациона нивоа $\pm y_1$, што је и приказано у боји на крајње десним хистограмима, док је SQNR повећан у поређењу са двобитним униформним квантизером без слојевитог адаптирања, што је и приказано у табели 4.2.2.2.

Коначно, упоређујемо тачност коју је постигао двобитни униформни квантизер у нашем повољном случају, означеном као случај 1, са сличним решењима упоредиве сложености NN модела након обуке, доступним у литератури. Тачност постигнута коришћењем нашег решења у пост-тренинг фази је очигледно већа у поређењу са онима представљеним у [40], [77] - [79]. У [78], је битска брзина само 1 бит/одмерак.

У [40] и [79], је коришћен двобитни униформни квантизер, са разликом у дефиницији величине корака квантовања ($\Delta = 2\, x_{\max}/(N-1)$) и у избору $\Re_g$. Приближнији случај за поређење је онај који је представљен у [77], где је представљен адаптивни двобитни униформни квантизер. Наш повољан дизајн квантизера, означен случајем 1, обезбедио је већу тачност QNN на скупу за валидацију, за 0.71 % и за 1 % у случају адаптације $\Re_g$ по слојевима. Истакнимо да у [77], тачност модела NN који је трениран тежинама у FP32 формату износи 96.86 %, док је тачност нашег NN модела пре компресије 98.1 %. Како је дубина наше NN већа од MLP из [77], очекивано је да је тачност нашег модела QNN већа. Међутим, пошто је број параметара за квантизацију у нашем моделу много већи у односу на модел из [77], већи број параметара модела пружа могућност да се тачност још више деградира, што још једном потврђује важаност избора $\Re_g$, односно анализе која је овде изложена.

Табела 4.2.2.3. Поређење тачности QNN модела.

	1-битни [78]	2-битни UQ [79]	2-битни UQ [40]	2-битни адаптирани UQ [77]	Наш случај 1 [21]	Наш слојевити случај 1 [21]
Тачност [%]	91.12	94.70	94.49	96.26	96.97	97.26

4.2.3 Квалитативни показатељи перформанси двобитног униформног квантизера и униформно квантоване неуронске мреже

За оптимизацију квалитативних показатеља двобитног униформног квантизера примењеног у QNN могу се специфицирати различити креитеријуми, као што је описано у раду [22]:

Критеријум 1: $x_{\max}^{\text{opt SQNR}_{\text{th}}^{\text{UQ}}} = \arg\max \text{SQNR}_{\text{th}}^{\text{UQ}}(x_{\max})$,

Критеријум 2: $x_{\max}^{\text{opt SQNR}_{\text{ex}}^{\text{UQ}}} = \arg\max \text{SQNR}_{\text{ex}}^{\text{UQ}}(x_{\max})$,

Критеријум 3: $x_{\max}^{\text{opt Acc}} = \arg\max \text{Acc}(x_{\max})$. (4.2.3.1)

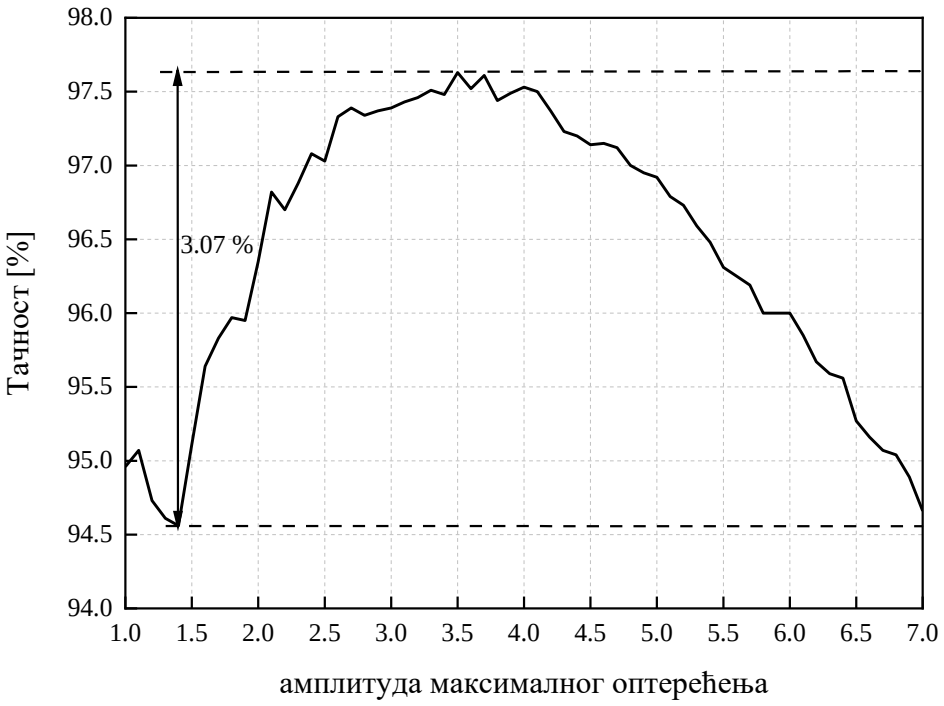
Резултат примене првог критеријума је јединствена вредност $x_{\max}$ ($x_{\max} = x_{\max}^{\text{opt SQNR}_{\text{th}}^{\text{UQ}}}$) за коју се постиже максимум теоријски утврђеног SQNR двобитног униформног квантизера за Лапласов извор. Применом другог критеријума одређујемо $x_{\max}$ ($x_{\max} = x_{\max}^{\text{opt SQNR}_{\text{ex}}^{\text{UQ}}}$) за који се постиже максимум експериментално постигнутог SQNR двобитног униформног квантизера. На крају, резултат примене последњег критеријума оптимизације је вредност $x_{\max}$ ($x_{\max} = x_{\max}^{\text{opt Acc}}$) за коју се постиже максимална тачност нашег модела QNN. Критеријум 1 даје $x_{\max}$ = 2.2 за највећу вредност теоријског SQNR од 7.06953 dB. Табела 4.2.3.1. показује да у овом случају добијамо најнижу тачност QNN међу посматраним критеријумима.

Табела 4.2.3.1. Резултати оптимизације применом три критеријума дефинисаних за двобитни UQ и поређење перформанси.

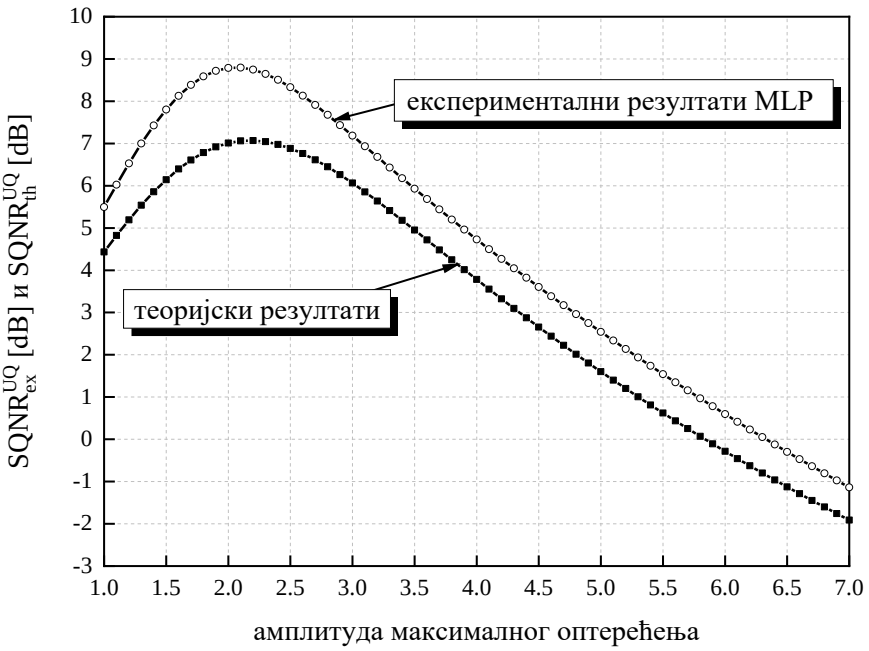
	Критеријум 1	*Критеријум* 2	*Критеријум* 3
$x_{\max}$	2.2	2.1	3.5
$\text{SQNR}_{\text{th}}^{\text{UQ}}$[dB]	**7.06953**	7.05978	4.95028
$\text{SQNR}_{\text{ex}}^{\text{UQ}}$[dB]	8.7465	**8.79551**	5.92983
Тачност [%]	96.7	96.82	**97.63**

Критеријум 2 даје сличну x_{max} вредност (x_{max} = 2.1), за који смо добили највећи експериментално утврђен SQNR. У овом случају примећујемо благи пораст тачности QNN, који је и даље далеко од максималне тачности (видети слику 4.2.3.1.).

Коначно, критеријум 3 даје вредност x_{max} (x_{max} = 3.5) која обезбеђује највећу тачност QNN од 97.63 % (видети слику 4.2.3.1.). Иако критеријум 3 пружа највећу тачност QNN, он такође пружа најниже теоријске и експерименталне вредности SQNR (видети слику 4.2.3.2.) и много шири грануларни регион у поређењу са претходним критеријумима. Ови резултати потврђују да x_{max} представља најважнији параметар по питању тачности QNN, који нема једнозначну везу са вредношћу оствареног SQNR за двобитни униформни квантизер.



Слика 4.2.3.1. Тачност двобитног униформног квантизера за модел QNN трениран на MNIST скупу, за опсег вредности амплитуде максималног оптерећења од x_{max} = 1 до $|x_{min}|$ = 7.063787.



Слика 4.2.3.2. Теоријски и експериментални SQNR за опсег вредности амплитуде максималног оптерећења од $x_{max} = 1$ до $|x_{min}| = 7.063787$.

У овом одељку смо показали да чак и када се „агресивна“ двобитна униформна квантизација користи за квантизацију тежина након тренинга, тачност разматраног модела NN може бити мало или незнатно деградирана, уколико се пажљиво изабере ширина грануларног региона квантизера. Интуитивно смо у иницијаном истраживању описаном у [21] исказали очекивање да ће избор ширине грануларног региона, $\Re_g$, имати велики утицај на тачност NN модела. Да бисмо потврдили интуицију, дизајнирали смо модел двобитног униформног квантизера за различите ширине $\Re_g$. Један интересантан закључак изведен и [21] и [22], односи се на чињеницу да избор x_{max} за који се постиже максималан SQNR нужно не подразумева и максималну тачност модела NN за тај исти x_{max}. Сличне анализе, а потенцијално и закључци, могли би се лако извести и за неке друге скупове података за које би се тренирала NN, пошто је добро познато да тежине у многим NN подлежу Лапласовој расподели, која је и претпостављена у нашим анализама.

4.2.4 Анализа утицаја избора амплитуде максималног оптерећења тробитног униформног квантизера на перформансе у пост-тренинг квантизацији

Овај одељак описује ефикасно решење за компресију тежина у неуронским мрежама користећи једноставан тробитни униформни квантизер [24]. Као што је насловљено, анализираће се избор кључног параметра разматраног квантизера (тј. амплитуде максималног оптерећења) и пружиће се детаљан преглед утицаја њеног избора на перформансе квантизације NN након обуке за MNIST скуп података, примарно. Ближе речено, анализираће се да ли је применом једноставног тробитног униформног квантизера могуће у великој мери сачувати тачност нашег NN модела, без обзира на избор његовог кључног параметра. Познато је да је код квантизације одмерака сигнала, одређивање овог кључног параметра од највеће важности, те је такође циљ одредити и степен његове важности код квантованих тежина NN.

Подсетимо, у одељку 4.2.2 је дата анализа перформанси примењеног униформног квантизера са битском брзином од R = 2 бит/одмерак за представљање тежина обученог вишеслојног перцептрона (MLP) уз параметризацију квантизера која омогућава значајно очување тачности. Како смо показали да избор амплитуде максималног оптерећења има велики утицај на тачност QNN, у овом одељку, који се и даље фокусира на нискобитне презентације, циљ је да утврдимо да ли се могу дати потпуно нови увиди и закључци у случају тробитне униформне квантизације, а за исти задатак класификације. Показали смо у 4.2.2 да је добром параметризацијом двобитног униформног квантизера, који је коришћен за компресију тежина код MLP модела, могуће сачувати тачност QNN, која је деградирана за нешто више од 1 %, док је број тежина у FP32 формату компримован 16 пута. Такође смо показали у 4.2.2 и да неодговарајући избор амплитуде максималног оптерећења може значајно деградирати тачност NN (више од 3.5 %) као и перформансе двобитне униформне квантизације примењене у фази после обуке на претпостављеном MLP моделу, обученом на MNIST скупу података. Како бисмо вредности квантованих тежина одржали што ближе

оригиналним, а циљајући барем упола смањену деградацију тачности од оне дате у одељку 4.2.2, у раду [24] смо истраживали да ли је уз далеко једноставнији и мање строжи избор амплитуде максималног оптерећења за тробитни униформни квантизер, постављени циљ и достижан.

Пошто је SQNR објективни индикатор перформанси квантизације, од посебног интереса је показати како избор кључног параметра, тј. амплитуде максималног оптерећења утиче и на SQNR и на тачност QNN. Један додатни бит за квантизацију тежина је важан за све слојеве, почевши од првог слоја (кључног за следеће слојеве пошто представља почетне тежине што је могуће тачније), до последњег слоја који директно одређује резултат класификације. Истакнимо да смо у одељку 4.2.2 извели и један занимљив закључак о адаптацији по слојевима, пошто смо приметили да расподеле тежина варирају у различитим слојевима MLP. Конкретно, приметили смо да на трећем MLP слоју постоји највеће одступање од Лапласове расподеле, те да се та расподела приближава униформној, што је посебно погодно за примену униформне квантизације.

У овој анализи ће се, као и у претходним одељцима, такође примењивати нормализација тежина NN. Наиме, методи нормализације тежина NN су већ потврдили добру емпиријску функционалност као нпр. у [47], при чему многи од њих имплицитно претпостављају да раподеле нормализованих NN параметара, пре свега тежина, имају нулту средњу вредност и јединичну варијансу. Имајући у виду да расподела тежина може блиско одговарати некој од добро познатих функција густине вероватноће, као што је Лапласова функција густине вероватноће [16], у овом одељку, као и у [45], [49], [77], [78], претпостављамо расподелу сличну Лапласовој за експерименталну расподелу тежина и Лапласову функцију густине вероватноће за теоријску расподелу тежина код процене перформанси тробитног униформног квантизера, који описујемо у наставку. Како би даље проширили анализу нумеричких и експерименталних резултата и извели коректне закључке, у раду [24] извели смо перформансну анализу тробитне униформне квантизације тежина одређених након обуке за CNN која је коришћена за класификацију MNIST скупа података

За симетричне расподеле сигнала или података, пожељни су симетрични квантизери са парним бројем нивоа квантизације[12] - [14], [17], [18], [30], [57], [58], [77], [80], [81]. Иако се може очекивати да је расподела амплитуда већине стварних сигнала или података асиметрична, не мора се нужно претпоставити да је пожељни квантизер такође асиметричан. За боље разумевање процеса квантизације који узима у обзир стварне, не нужно симетричне податке, у овом одељку описујемо симетрични тробитни униформни квантизер и анализирамо његове перформансе за Лапласову функцију густине вероватноће. У наредном одељку описујемо његово прилагођавање стварним подацима, односно нормализацију тежина претходно обучене NN. Пошто нискобитна квантизација може довести до значајног смањења SQNR, описаћемо теоријске и експерименталне сценарије за тробитну униформну квантизацију FP32 тежина два истренирана NN модела, MLP и CNN. Као и у [21], претпостављамо расподелу сличну Лапласовој за експерименталну расподелу тежина и Лапласову функцију густине вероватноће за теоријски представу тежина како би проценили експерименталне и теоријске перформансе тробитног униформног квантизера, чији дизајн описујемо у наставку.

Позваћемо се на раније поменуте полазне поставке о униформној квантизацији и симетричном униформном квантизеру са N нивоа, грануларном региону $\Re_g$ и региону прекорачења $\Re_o$. За дату брзину протока од R = 3 бит/одмерак са симетричним тробитним униформним квантизером, $\Re_g$ је подељен на $N = 2R = 8$ дисјунктних, ограничених у ширини и једнаких грануларних ћелија (видети слику 4.2.4.1.).

Тако је i-та грануларна ћелија тробитног униформног квантизера дефинисана као:

$$\Re_i = \{x \mid x \in [-x_{\max}, x_{\max}], Q_N(x) = y_i\}, \Re_i \cap \Re_j = \varnothing,\ i \neq j\,. \qquad (4.2.4.1)$$

$\{\Re_i\}_{i=-N/2}^{-1}$ и $\{\Re_i\}_{i=1}^{N/2}$ означавају грануларне ћелије у негативним и позитивним амплитудским регионима, који су симетрични. Ова симетрија важи и за квантизационе ћелије у региону прекорачења, односно за пар ћелија неограничене ширине у региону прекорачења $\Re_o$, који је дефинисан као:

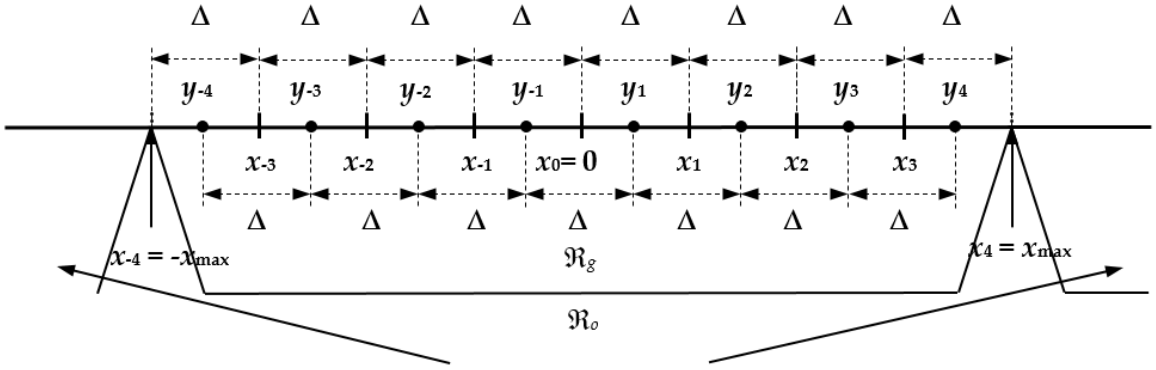
$$\Re_0 = \{x \mid x \notin [-x_{\max}, x_{\max}], Q_N(x) = y_{N/2},\ x > 0 \vee Q_N(x) = y_{-N/2},\ x < 0\}. \qquad (4.2.4.2)$$

Модел униформног квантизера подразумева да су репрезентациони нивои центрирани на средини одговарајућих квантизационих ћелија, а кодна књига $Y \equiv \{y_{-N/2}, \ldots, y_{-1}, y_1, \ldots, y_{N/2}\} \subset \mathbb{R}$ садржи N репрезентационих нивоа, означених са y_i. Параметри симетричног тробитног униформног квантизера могу се одредити из:

а) $\Delta = 2x_{\max}/N = 2x_{\max}/4$,

б) $x_i = i\,\Delta,\ x_{-i} = -x_i,\ i \in \{0,1,2,3,4\}$,

в) $y_i = (x_{i-1} + x_i)/2,\ y_{-i} = -y_i,\ i \in \{1,2,3,4\}$.

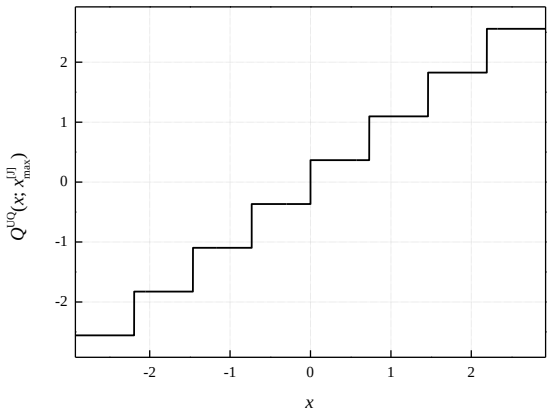


Слика 4.2.4.1. Грануларни регион $\Re_g$ и регион прекорачења $\Re_o$ симетричног тробитног униформног квантизера.

Будући да важи да $x_{\max} = x_4$, где $x_{\max}$ означава амплитуду максималног оптерећења тробитног униформног квантизера, из правила *а*) - *в*) се може закључити да $x_{\max}$ у потпуности одређује ширину ћелије Δ, границе ћелије x_i, $i \in \{0,1,2,3,4\}$ и репрезентационе нивое y_i, $i \in \{0,1,2,3,4\}$ тробитног униформног квантизера (видети сл. 4.2.4.1.). Сходно томе, као што смо већ поменули, $x_{\max}$ је кључни параметар униформног квантизера. За дату амплитуду максималног оптерећења $x_{\max}$, униформни квантизер је такође добро описан својом преносном карактеристиком, која ја за симетрични тробитни униформног квантизера дата као:

$$Q^{\mathrm{UQ}}\left(x;x_{\max}\right)=\begin{cases}\operatorname{sgn}(x)\left(\left\lfloor |x|/\Delta \right\rfloor+1/2\right)\Delta, & |x|\le x_{\max}\\ \operatorname{sgn}(x)\left(x_{\max}-\Delta/2\right), & |x|>x_{\max}\end{cases}, \tag{4.2.4.3}$$

где је ознака N, која означава број нивоа квантовања, изостављена због једноставности. За $x_{\max} = x_{\max}$ [J], дата је илустрација на слици 4.2.4.2, а означена је са $Q^{\mathrm{UQ}}(x; x_{\max}^{[\mathrm{J}]})$.



Слика 4.2.4.2. Преносна карактеристика симетричног тробитног UQ, $Q^{\mathrm{UQ}}(x; x_{\max}^{[\mathrm{J}]})$.

За дизајн тробитног униформног квантизера у обради сигнала је битан максимални SQNR, односно минимална укупна дисторзија, јер $x_{\max}$ тада има оптималну вредност. Да бисмо одредили укупну дисторзију нашег тробитног униформног квантизера, $D^{\mathrm{UQ}}=D_{\mathrm{g}}^{\mathrm{UQ}}+D_{\mathrm{o}}^{\mathrm{UQ}}$ кренућемо од основне дефиниције дисторзије, датом изразом (1) [2]:

$$D_{\mathrm{g}}^{\mathrm{UQ}}=2\sum_{i=1}^{4}\int_{x_{i-1}}^{x_i}\left(x-y_i\right)^2\,p(x)dx, \tag{4.2.4.4}$$

$$D_{\mathrm{o}}^{\mathrm{UQ}}=2\int_{x_4}^{\infty}\left(x-y_4\right)^2\,p(x)dx. \tag{4.2.4.5}$$

Уводимо да је $x_5 = \infty$, где x_5 означава горњу границу интеграла у изразу (4.2.4.5). Подсетимо се да су границе квантизационих ћелија и репрезентациони нивои директно одређени амплитудом максималног оптерећења, $x_{\max}$. Сходно томе, формуле (4.2.4.4) и (4.2.4.5) указује на то да се $x_{\max}$ имплицитно јавља не само у границама интеграла, већ и у интеграндима. Због тога, упркос једноставности модела униформног

квантизера, аналитичко одређивање $x_{\max}$ из услова минималне MSE дисторзије је и даље тешко, или чак немогуће за многе функцију густине вероватноће. Вредност $x_{\max}$ коју је одредио *Jayant* [2] је резултат нумеричке оптимизације дисторзије. *Hui* је аналитички дошао до једначине за $x_{\max}$ симетричног асимптотски оптималног униформног квантизера са N нивоа, који је дизајниран за велике битске брзине и улазни сигнал са Лапласовом функцију густине вероватноће нулте средње вредности и јединичне варијансе [62], што за $N = 2R = 8$ даје:

$$x_{\max}^{[\mathrm{H}]} = \sqrt{2}\ln(8), \tag{4.2.4.6}$$

Као што смо раније истакли, претпостављамо неограничену Лапласову функцију густине вероватноће нулте средње вредности и варијансе $\sigma^2 = 1$ за моделовање улазних података квантизера у теоријској анализи описаног тробитног униформног квантизера:

$$p(x) = \frac{1}{\sqrt{2}\sigma}\exp\left\{-\frac{\sqrt{2}x}{\sigma}\right\}. \tag{4.2.4.7}$$

Укупна дисторзија описаног симетричног тробитног униформног квантизера је:

$$D^{\mathrm{UQ}} = 2\sum_{i=1}^{5}\int_{x_{i-1}}^{x_i} x^2 p(x)dx - 4\left(\sum_{i=1}^{4} y_i \int_{x_{i-1}}^{x_i} xp(x)dx + y_4\int_{x_4}^{\infty} xp(x)dx\right) +,$$

$$+2\left(\sum_{i=1}^{4} {y_i}^2 \int_{x_{i-1}}^{x_i} p(x)dx + {y_4}^2\int_{x_4}^{\infty} p(x)dx\right), \tag{4.2.4.8}$$

$$D^{\mathrm{UQ}} = \Upsilon^{\mathrm{I}} - 4\left(\sum_{i=1}^{4} y_i \Upsilon_i^{\mathrm{II}} + y_4\Upsilon_5^{\mathrm{II}}\right) + 2\left(\sum_{i=1}^{4} {y_i}^2 \Upsilon_i^{\mathrm{III}} + {y_4}^2\Upsilon_5^{\mathrm{III}}\right). \tag{4.2.4.9}$$

За претпостављену функцију густине вероватноће дате са (4.2.4.7), даље изводимо:

$$\Upsilon^{\mathrm{I}} = 2\sum_{i=1}^{5}\int_{x_{i-1}}^{x_i} x^2 p(x)dx = \sigma^2, \tag{4.2.4.10}$$

$$\Upsilon_i^{\mathrm{II}} = \int_{x_{i-1}}^{x_i} xp(x)dx = \frac{1}{2}\left[\left(x_{i-1} + \frac{\sigma}{\sqrt{2}}\right)\exp\left\{-\frac{\sqrt{2}x_{i-1}}{\sigma}\right\} - \left(x_i + \frac{\sigma}{\sqrt{2}}\right)\exp\left\{-\frac{\sqrt{2}x_i}{\sigma}\right\}\right], i = 1,...,5, \quad (4.2.4.11)$$

$$\Upsilon_i^{\mathrm{III}} = \int_{x_{i-1}}^{x_i} p(x)dx = \frac{1}{2}\left[\exp\left\{-\frac{\sqrt{2}x_{i-1}}{\sigma}\right\} - \exp\left\{-\frac{\sqrt{2}x_i}{\sigma}\right\}\right], i = 1,...,5. \quad (4.2.4.12)$$

На крају, заменом (4.2.4.10), (4.2.4.11) и (4.2.4.12) у (4.2.4.9) добијамо:

$$D^{\mathrm{UQ}} = \sigma^2 + \left(\frac{x_{\max}}{8}\right)^2 - \sqrt{2}\sigma\frac{x_{\max}}{8}\left[1 + 2\sum_{i=1}^{3}\exp\left\{-i\frac{\sqrt{2}x_{\max}}{4\sigma}\right\}\right]. \quad (4.2.4.13)$$

Слично, као и у бројним радовима у којима је примењена квантизација (нпр. у [12] - [14], [30], [39], [57], [58], [62]), заинтересовани смо да спроведемо анализу за случај прилагођене варијансе, где је пројектована, уједно и варијанса сигнала или података, доведених на улаз квантизера и обично износи $\sigma^2 = 1$, што нас доводи до нове формуле у затвореном облику за дисторзију симетричног тробитног униформног квантизера:

$$D^{\mathrm{UQ}}\Big|_{\sigma^2=1} = 1 + \left(\frac{x_{\max}}{8}\right)^2 - \sqrt{2}\frac{x_{\max}}{8}\left(1 + 2\sum_{i=1}^{3}\exp\left\{-i\frac{\sqrt{2}x_{\max}}{4}\right\}\right). \quad (4.2.4.14)$$

На крају овог одељка дефинишемо теоријски SQNR као:

$$\mathrm{SQNR}_{\mathrm{th}}^{\mathrm{UQ}} = 10\log_{10}\left(\frac{\sigma^2}{D^{\mathrm{UQ}}}\right), \quad (4.2.4.15)$$

који ћемо срачунати за $\sigma^2 = 1$ и поредити са експериментално утврђеним $\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{UQ}}$.

4.2.5 Примена тробитног униформног квантизера у пост-тренинг квантизацији

Као што је раније поменуто, компресија тежина NN у различитим сценаријима QNN подразумева смањење броја репрезентација квантованих тежина уз пожељно најмањи могући губитак тачности QNN. Код компресије са губицима, где се оригиналне улазне вредности тежина NN не могу идеално повратити након примене компресије или квантизације, уводи се неповратна грешка услед квантизације, која може имати негативан утицај на тачност QNN. Неки QNN примењују квантизацију тежина у фази обуке и у одређеној мери обезбеђују опоравак тачности QNN кроз фазу поновне обуке. Како је у овом поглављу квантизација коришћена у фази након обуке, у наставку описујемо саму пост-тренинг процедуру, која укључује три секвенцијалне операције примењене на тежине NN: нормализацију, квантизацију и денормализацију.

Нормализација тежина је прва операција, примењена на све тежине представљене једнодимензионалним вектором $\hat{\boldsymbol{W}}=\{w_j\}_{j=1,2,\ldots,W}$ од W елемената, где је W = 669 706 укупан број тежина за наш MLP модел. Свака тежина се нормализује, чиме се формира вектор $\hat{\boldsymbol{W}}^{\mathrm{N}} = \{w_j^{\mathrm{N}}\}_{j=1,2,\ldots,W}$

$$w_j^{\mathrm{N}} = \frac{w_j - \mathrm{mean}(\hat{\boldsymbol{W}})}{\mathrm{std}(\hat{\boldsymbol{W}})}, \tag{4.2.5.1}$$

где су mean($\hat{\boldsymbol{W}}$) и std($\hat{\boldsymbol{W}}$), средња вредност и стандардна девијација вектора $\hat{\boldsymbol{W}}$, респективно. Укратко, тежине представљене у FP32 формату су нормализоване тако да имају нулту средњу вредност и јединичну варијансу пре него што доведу на улаз квантизера. Када се тежине нормализују, формира се дискретна расподела нормализованих тежина са $w_{\min}$ и $w_{\max}$ који означавају минималну и максималну вредност нормализованих тежина, при чему је ознака N која означава процедуру нормализације овде изостављена због једноставности.

Други корак у процедури пост-тренинг квантизације је примена тробитне униформне квантизације на нормализоване тежине. Поступком квантизације, све

нормализоване тежине су подељене у два комплементарна скупа: скуп тежина које припадају грануларном региону [$-w_{\text{supp}}$, w_{supp}] и скуп тежина које му не припадају. У квантизацији је пожељано да грануларни регион [$-w_{\text{supp}}$, w_{supp}] обухвата веома велики број тежина. За дату вредност w_{supp}, вршимо униформну квантизацију нормализованих тежина коришћењем:

$$w_j^{\text{UQN}} = Q_{w_j^{\text{N}}}^{\text{UQ}}\left(w_j^{\text{N}}; w_{\text{supp}}\right) = \begin{cases} \text{sgn}(w_j^{\text{N}})\left(\left\lfloor \dfrac{4\left|w_j^{\text{N}}\right|}{w_{\text{supp}}} \right\rfloor + \dfrac{1}{2}\right)\dfrac{w_{\text{supp}}}{4}, & \left|w_j^{\text{N}}\right| \le w_{\text{supp}} \\ \text{sgn}(w_j^{\text{N}}) \cdot \dfrac{7w_{\text{supp}}}{8}, & \left|w_j^{\text{N}}\right| > w_{\text{supp}} \end{cases}, \qquad (4.2.5.2)$$

формирајући $\hat{\boldsymbol{W}}^{\text{UQN}} = \{w_j^{\text{UQN}}\}_{j=1,2,\ldots,W}$, са 8 различитих вредности за тробитни униформни квантизер.

Алгоритам 4.2.5.1. - Компресија тежина помоћу квантизације након обуке користећи симетрични тробитни униформни квантизер

Нотација: w_j – тежина обучене NN, w_j^{N} – нормализована тежина, w_j^{UQN} – униформно квантована нормализована тежина, w_j^{UQ} – униформно квантована тежина
Улаз: $\hat{\boldsymbol{W}} = \{w_j\}_{j=1,2,\ldots,W}$, тежине представљене у FP32 формату
Излаз: $\hat{\boldsymbol{W}}^{\text{UQ}}=\{w_j^{\text{UQ}}\}_{j=1,2,\ldots,W}$, униформно квантоване тежине, $\text{SQNR}_{\text{th}}^{\text{UQ}}$, $\text{SQNR}_{\text{ex}}^{\text{UQ}}$, тачност

учитавање почетних унапред обучених тежина $\hat{\boldsymbol{W}} = \{w_j\}_{j=1,2,\ldots,W}$
израчунавање средње вредности mean($\hat{\boldsymbol{W}}$) и стандардне девијације std($\hat{\boldsymbol{W}}$)
нормализација тежина коришћењем (4.2.5.1)
формирање $\hat{\boldsymbol{W}}^{\text{N}}=\{w_j^{\text{N}}\}_{j=1,2,\ldots,W}$
$w_{\min} \leftarrow$ минимална вредност нормализованих тежина од $\hat{\boldsymbol{W}}^{\text{N}}$
$w_{\max} \leftarrow$ максимална вредност нормализованих тежина од $\hat{\boldsymbol{W}}^{\text{N}}$
Избор w_{supp}// $w_{\text{supp}} \leftarrow |w_{\min}|$ или $w_{\max}$ или $x_{\max}^{[J]}$ или $x_{\max}^{[H]}$ или неке друге дате вредности
$j \leftarrow 1$
while $j \le W$ **do**
квантизација w_j^{N} коришћењем (4.2.5.2)
end while
денормализација квантованих тежина
формирање $\hat{\boldsymbol{W}}^{\text{UQ}}$
позивање (4.2.5.3) за израчунавање $D_{\text{ex}}^{\text{UQ}}$
израчунавање $\text{SQNR}_{\text{ex}}^{\text{UQ}}$ коришћењем (4.2.5.4)
16. израчунавање $\text{SQNR}_{\text{th}}^{\text{UQ}}$ коришћењем (4.2.4.15)
процена тачности
$\hat{\boldsymbol{W}}^{\text{UQ}}$, $\text{SQNR}_{\text{th}}^{\text{UQ}}$, $\text{SQNR}_{\text{ex}}^{\text{UQ}}$, тачност

Последњи корак у описаној процедури пост-тренинг квантизације је денормализација. Денормализација квантованих тежина се врши након њихове квантизације тако да се квантоване тежине које се денормализују, $\hat{\boldsymbol{W}}^{\mathrm{UQ}}=\{w_j^{\mathrm{UQ}}\}_{j=1,2,\ldots,W}$, могу вратити у првобитни опсег вредности и учитати у QNN модел.

На крају, процену перформанси тробитног униформног квантизера можемо одредити помоћу дисторзије и SQNR као:

$$D_{\mathrm{ex}}^{\mathrm{UQ}}=\frac{1}{W}\left\|\hat{\boldsymbol{W}}-\hat{\boldsymbol{W}}^{\mathrm{UQ}}\right\|_2^2=\frac{1}{W}\left(\hat{\boldsymbol{W}}-\hat{\boldsymbol{W}}^{\mathrm{UQ}}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{W}}-\hat{\boldsymbol{W}}^{\mathrm{UQ}}\right)=\frac{1}{W}\sum_{j=1}^{W}\left(w_j-w_j^{\mathrm{UQ}}\right)^2, \qquad (4.2.5.3)$$

$$\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{UQ}}=10\log_{10}\left(\frac{\frac{1}{W}\sum_{j=1}^{W}w_j^2}{D_{\mathrm{ex}}^{\mathrm{UQ}}}\right). \qquad (4.2.5.4)$$

4.2.6 Експериментални резултати и анализа примене тробитног униформног квантизера за компресију MLP и CNN модела

Анализа перформанси квантизације тежина QNN у пост-тренинг фази може се посматрати кроз два аспекта: тачност QNN и SQNR униформног квантизера, примењеног за компресију тежина. За оба модела NN, MLP и CNN, које су обучене на MNIST скупу података, у раду [24] одредили смо тачност QNN и SQNR за различите вредности амплитуде максималног оптерећења и за битску брзину R = 3 бит/одмерак. Као што је раније поменуто, са тежинама у FP32 формату, тачност MLP модела на MNIST скупу података за валидацију пре примене квантизације износи 98.1 %, што је референтна тачност која се користи за процену деградације тачности NN модела у овој дисертацији. Друга референтна тачка је тачност QNN од 96.97 % представљена у одељку 4.2.2 тј. у раду [21], која је постигнута употребом двобитног униформног квантизера за квантизацију идентичних тежина, као и са идентичном MLP архитектуром. Наиме, друга референтна тачка ће нам дати увид у разлике у утицају $\Re_g$ на тачност QNN када се користи 1 бит/одмерак више за компресију тежина. Интуитивно се може претпоставити да ће за већу битску брзину (R = 3 бит/одмерак), утицај амплитуде максималног оптерећања квантизера бити мањи, а да бисмо потврдили ову премису, у раду [24] смо анализирали тачности QNN и SQNR за више различитих избора $\Re_g$ тробитног униформног квантизера.

Табела 4.2.6.1. приказује раподелу тежина трениране NN и перформансе описаног MLP модела пре и након квантизације тежина, за четири независна тренинга истог MLP модела на MNIST скупу података. Може се приметити да тежине потпадају у скоро исти опсег вредности за сваки тренинг, невезано од тога да ли је нормализација примењена или не. Такође, NN модел постиже сличну тачност са мањим варијацијама током поновљених тренинга. Додатно, постигнуте SQNR вредности (за случај где је амплитуда максималног оптерећења специфицирана са (4.2.4.6)) су веома блиске за различите једнодимензионалне векторе тежина са динамиком од приближно 0.04 dB.

Табела 4.2.6.1. Анализа статистике MLP тежина: амплитудска динамика пре/после нормализације на нулту средњу вредност и јединичну варијансу, тачност пре/после квантизације тробитним униформним квантизером и SQNR.

име фајла тежина	min тежине FP32	max тежине FP32	min тежине норм. FP32	max тежине норм. FP32	Тачност FP32 [%]	Тачност $UQ^{[H]}$ [%]	SQNR $UQ^{[H]}$ [dB]
MNIST_ тежине	-0.5185	0.3447	-7.0638	4.8371	**98.1**	97.66	12.9455
MNIST_тежине_1	-0.4948	0.3594	-6.6845	5.0123	98.28	98.17	12.9623
MNIST_тежине_2	-0.5122	0.3533	-6.960	4.948	98.18	97.9	12.9021
MNIST_тежине_3	-0.5096	0.3624	-6.9375	5.0808	98.3	97.92	12.9204

Табела 4.2.6.2. даје приказ тачности QNN и SQNR тробитног униформног квантизера, остварену за различите изборе $\Re_g$, битску брзину од R = 3 бит/одмерак и раније описани MLP. Заједно са експериментално оствареним SQNR вредностима (одређеним коришћењем (4.2.5.3) и (4.2.5.4)), одређен је и теоријски SQNR користећи (4.2.4.14) и (4.2.4.15) за сваки од $\Re_g$ избора, како бисмо проценили разлике између теоријског и експерименталног SQNR.

Табела 4.2.6.2. SQNR и тачност QNN за MLP модел код различитих тробитних униформних квантизера.

$w_{min} = -7.063787$, $w_{max} = 4.8371024$, $x_{max}^{[H]} = 2.9408$, $x_{max}^{[J]} = 2.9236$	**Случај 1** $\Re_g$ $[-w_{max}, w_{max}]$	**Случај 2** $\Re_g$ $[w_{min}, -w_{min}]$	**Случај 3** $\Re_g$ $[-x_{max}^{[H]}, x_{max}^{[H]}]$	**Случај 4** $\Re_g$ $[-x_{max}^{[J]}, x_{max}^{[J]}]$
$SQNR_{ex}^{UQ}$[dB]	9.2302	5.9005	12.9455	**12.9766**
$SQNR_{th}^{UQ}$[dB]	8.6901	5.1273	11.4414	**11.4419**
Тачност [%]	**97.85**	**97.85**	97.66	97.65
Припадност $\Re_g$ [%]	99.988	100	99.403	99.381

Табела 4.2.6.3. SQNR и тачност QNN код различитих двобитних униформних квантизера.

$w_{min} = -7.063787$, $w_{max} = 4.8371024$, $x_{max}^{[H]} = 1.9605$, $x_{max}^{[J]} = 2.1748$	**Случај 1**	**Случај 2**	**Случај 3**	**Случај 4**
	$\Re_g$	$\Re_g$	$\Re_g$	$\Re_g$
	$[-x_{max}, x_{max}]$	$[x_{min}, -x_{min}]$	$[-x_{max}^{[H]}, x_{max}^{[H]}]$	$[-x_{max}^{[J]}, x_{max}^{[J]}]$
$SQNR_{ex}^{UQ}$[dB]	2.8821	-1.2402	**8.7676**	8.7639
$SQNR_{th}^{UQ}$[dB]	1.9360	-2.0066	6.9787	**7.0707**
Тачност [%]	**96.97**	94.58	96.34	96.74
Припадност $\Re_g$ [%]	99.988	100	94.787	96.691

Избори 1 и 2 зависе од расподеле нормализованих тежина NN модела, тачније од минималне и максималне вредности тежине (w_{min} и w_{max}), док је у изборима 3 и 4, $\Re_g$ специфициран добро познатим оптималним и асимптотски оптималним вредностима за претпостављену битску брзину.

Табела 4.2.6.3. даје SQNR и тачност QNN постигнуту применом двобитног униформног квантизера, где је $\Re_g$ дефинисан по истим принципима као у табели 4.2.6.2. а идентичне тежине MLP модела су квантоване (овде приказана табела из одељка 4.2.2, тј. из [21] ради лакшег поређења). Конкретно, избор 1 за $\Re_g$ у табели 4.2.6.2. поклапа се са случајем 1 у табели 4.2.6.3. итд.

Табела 4.2.6.3. је овде дата у компаративне сврхе, како би нам помогла у доношењу закључака о различитом утицају избора $\Re_g$ на SQNR и тачност QNN за разматрани MLP и различите ниске битске брзине.

У [21] је истакнуто да избор 2, односно случај 2, а као што се може приметити из табеле 4.2.6.3., није најпогоднији избор у погледу тачности QNN и SQNR. Зато ћемо у наставку анализирати резултате за MLP и тробитни униформни квантизер са референцирањем на двобитни униформни квантизер из одељка 4.2.2, тј. из рада [21].

Избор 1 дефинише $\Re_g$ униформног квантизера као [-w_{max}, w_{max}], што значи да је $\Re_g$ симетрично одређен према апсолутној максималној вредности нормализоване тежине. У нашем случају са MLP моделом, важи $|w_{max}| = w_{max}$ ($w_{max} \geq 0$), тако да поједностављујемо нотацију. Из табела 4.2.6.2. и 4.2.6.3. може се приметити да је са тако дефинисаним $\Re_g$, проценат нормализованих тежина унутар $\Re_g$ износи 99.988 %.

Избор 2 дефинише $\Re_g$ у опсегу [w_{min}, $-w_{min}$], што значи да је описани квантизер симетрично одређен према апсолутној минималној вредности нормализоване тежине $|w_{min}|$. У нашем случају, важи $|w_{min}| = -w_{min}$ ($w_{min} \leq 0$). $\Re_g$ дефинисан у избору 2 укључује све одмерке нормализованих тежина, што значи да ниједана од вредности нормализованих тежина не припада области прекорачења квантизера.

Из табеле 4.2.6.2. може се приметити да иако се SQNR вредности за изборе 1 и 2 у великој мери разликују, тачност QNN је идентична. За исте изборе 1 и 2 из табеле 4.2.6.3. се може запазити да са двобитним униформним квантизером, QNN показује значајне варијације перформанси и у оствареној SQNR вредности и тачности QNN, где су оба перформансна параметра доминантно одређена избором $\Re_g$.

Сличан је случај са изборима 3 и 4, где $x_{max}^{[H]}$ и $x_{max}^{[J]}$ имају релативно блиске вредности. SQNR вредности постигнуте за изборе 3 и 4, а у табелама 4.2.6.2. и 4.2.6.3. имају блиске вредности, што није случај са тачношћу QNN за уочена два избора 3 и 4.

Анализом тачности QNN за издвојена четири случаја избора $\Re_g$ тробитног униформног квантизера можемо уочити да динамика тачности QNN, дефинисана као разлика између највеће и најмање постигнуте тачности QNN, износи 0.2 %, док за идентична четири случаја избора $\Re_g$ и двобитног униформног квантизера ова динамика постигнуте тачности QNN износи 2.39 %.

Поређењем SQNR вредности постигнутих за тробитни униформни квантизер из табеле 4.2.6.2. за изборе 2 и 4, може се уочити велика динамичка SQNR вредности, која за изборе 2 и 4 износи 6.3146 dB, док разлика у тачности QNN за ове изборе $\Re_g$ износи само 0.2 %. Ово даље указује да велика разлика у постигнутим SQNR вредности

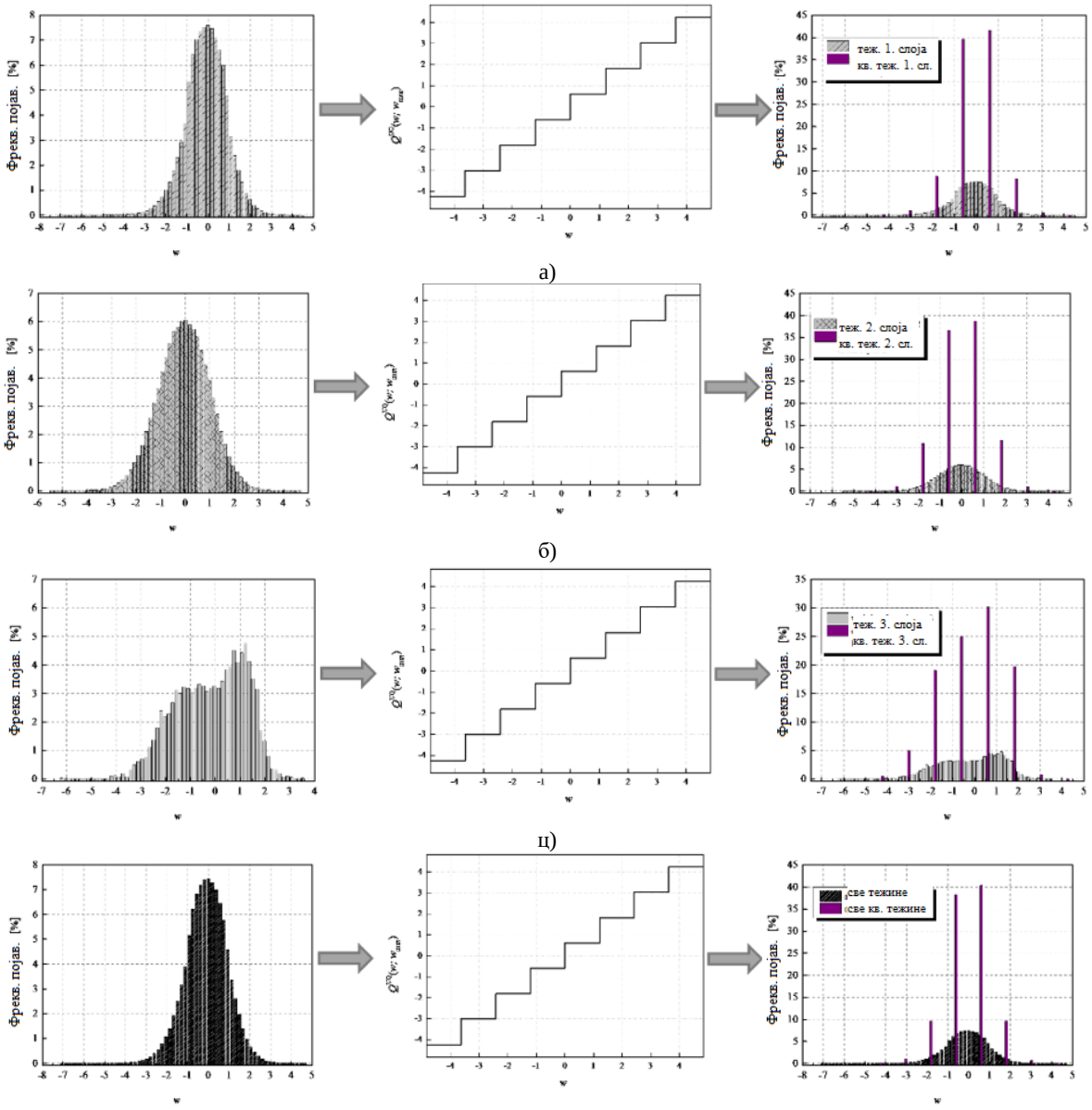
тробитног униформног квантизера не мора нужно да резултира великом разликом у тачности QNN, што је закључак који важи и за двобитни униформни квантизер дат у одељку 4.2.2, а истакнут је у раду [21].

Максимални теоријски и експериментални SQNR се постиже за избор 4, где је $\Re_g$ оптимално дефинисан са $x_{\max}^{[J]}$, док се максимална тачност QNN остварује изборима 1 и 2 (погледати подебљане вредности у табелама).

Компаративним приказом тачности QNN долазимо до закључка да избор $\Re_g$ доминантно одређује тачност QNN за случај битске брзине од 2 бит/одмерак, где имамо доступна само 4 репрезентациона нивоа. Удвостручавањем броја репрезентационих нивоа на 8, тј. за битску брзину од R = 3 бит/одмерак, утицај $\Re_g$ на тачност QNN се значајно смањује.

Да бисмо анализирали расподелу нормализованих тежина на различитим репрезентационим нивоима тробитног униформног квантизера, издвојили смо нормализоване хистограме нормализованих тежина појединачних слојева, као и заједнички нормализовани хистограм свих нормализованих тежина MLP модела, пре и после компресије/квантизације.

На први поглед, упоређивањем избора 1 и 2 са изборима 3 и 4, може се приметити да у прва два избора доминантно користимо само 4 од 8 доступних репрезентационих нивоа униформног квантизера, што је посебно уочљиво на нормализованом хистограму који представља све нормализоване тежине модела QNN за избор 2. Ово запажање нас наводи на закључак да у изборима 1 и 2 користимо непотребно широки $\Re_g$, што објашњава ниже вредности SQNR постигнуте за прва два избора, посебно у избору 2, где је $\Re_g$ интервал одређен као $[-7.063787, 7.063787]$. Истовремено, шири $\Re_g$ позитивно доприноси тачности QNN која је већи за прва два посматрана избора у поређењу са изборима 3 и 4.

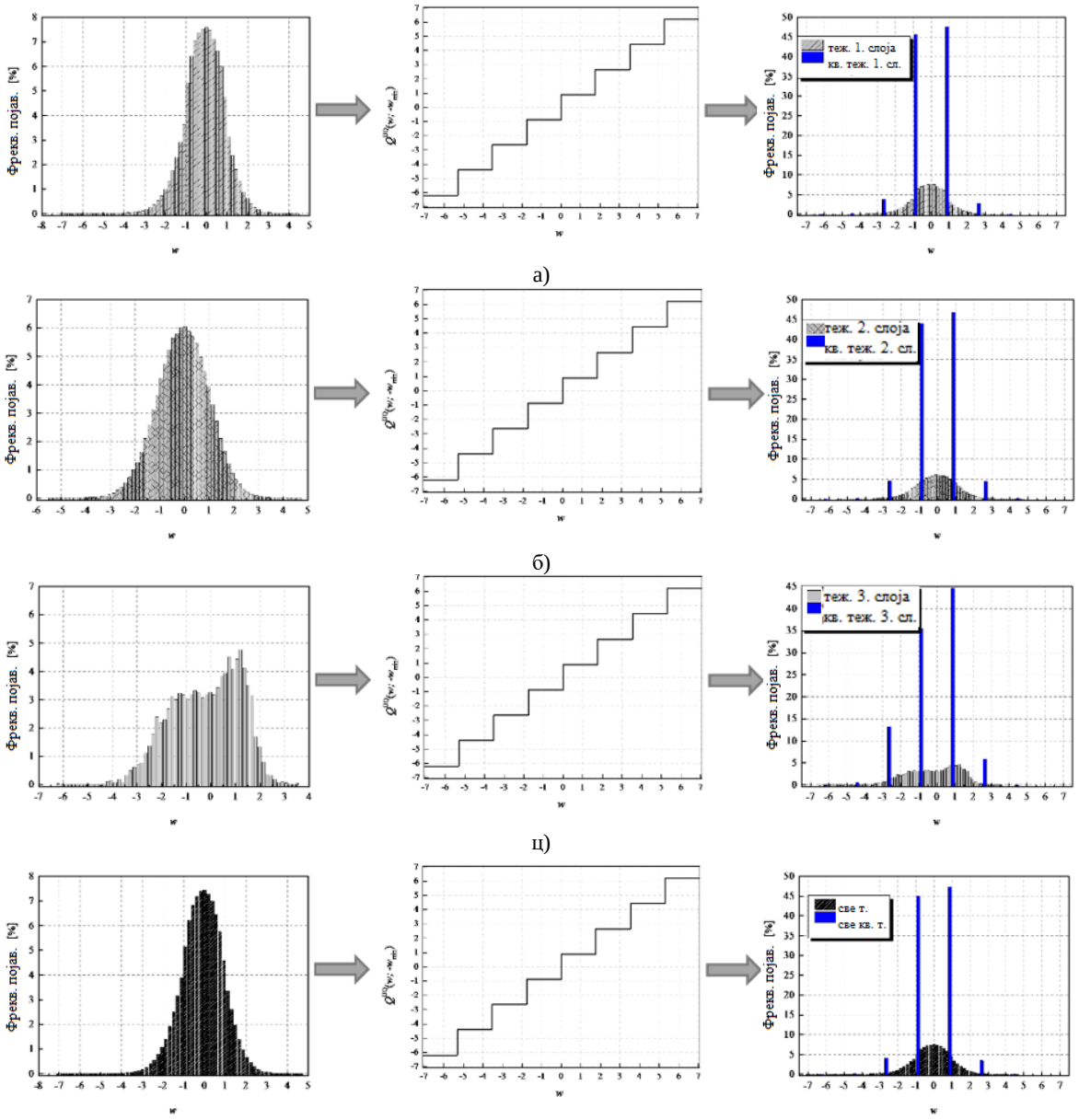


Слика 4.2.6.1 Нормализовани хистограм за MLP модел трениран на MNIST скупу са нормализованим FP32 тежинама за а) слој 1, б) слој 2, ц) слој 3 д) све слојеве;

Преносна карактеристика тробитног униформног квантизера за $\Re_g$ избор 1;

Нормализовани хистограм униформно квантованих нормализованих тежина за

а) слој 1, б) слој 2, ц) слој 3, д) све слојеве MLP.

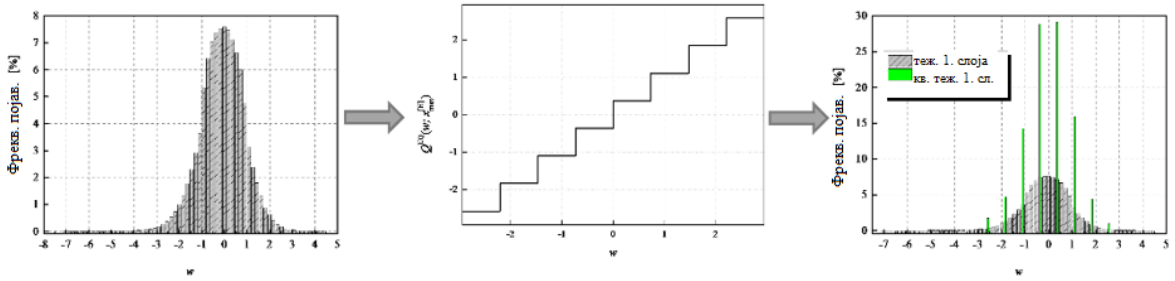


Слика 4.2.6.2. Нормализовани хистограм за MLP модел трениран на MNIST скупу са нормализованим FP32 тежинама за а) слој 1, б) слој 2, ц) слој 3 д) све слојеве;

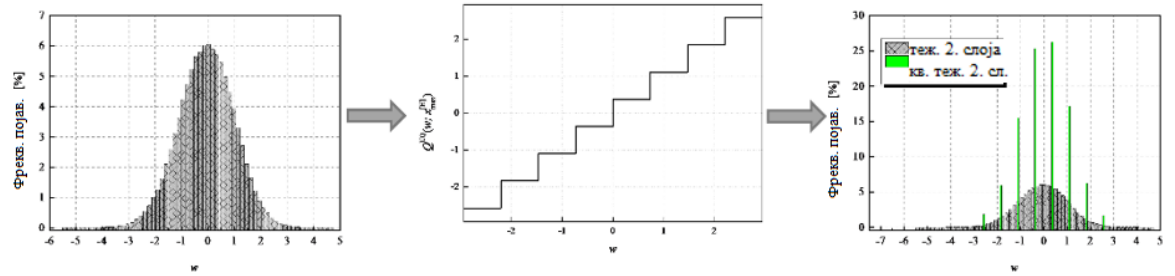
Преносна карактеристика тробитног униформног квантизера за $\Re_g$ избор 2;

Нормализовани хистограм униформно квантованих нормализованих тежина за

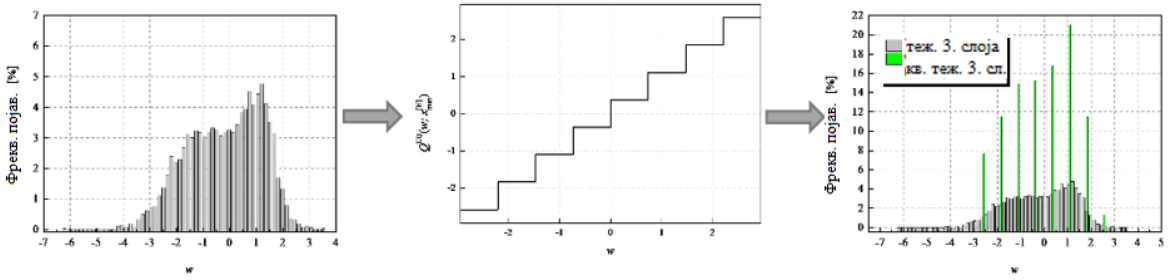
а) слој 1, б) слој 2, ц) слој 3, д) све слојеве MLP.



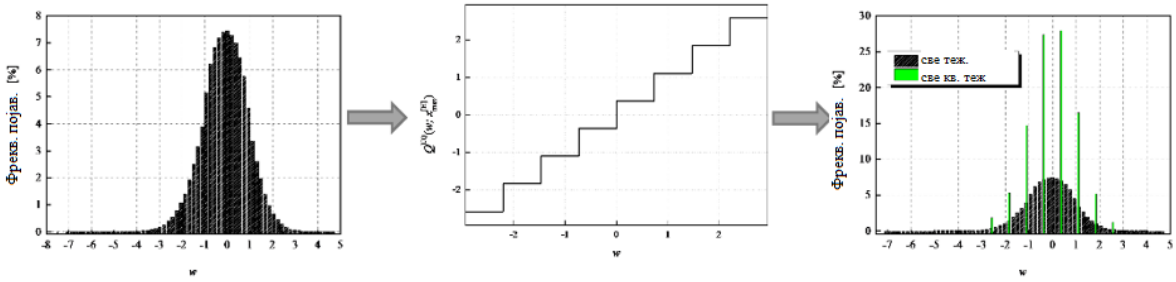
а)



б)



ц)



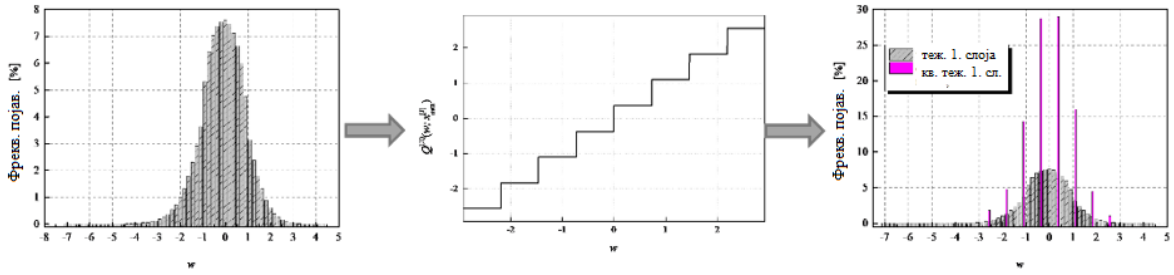
д)

Слика 4.2.6.3. Нормализовани хистограм за MLP модел трениран на MNIST скупу са нормализованим FP32 тежинама за а) слој 1, б) слој 2, ц) слој 3 д) све слојеве;

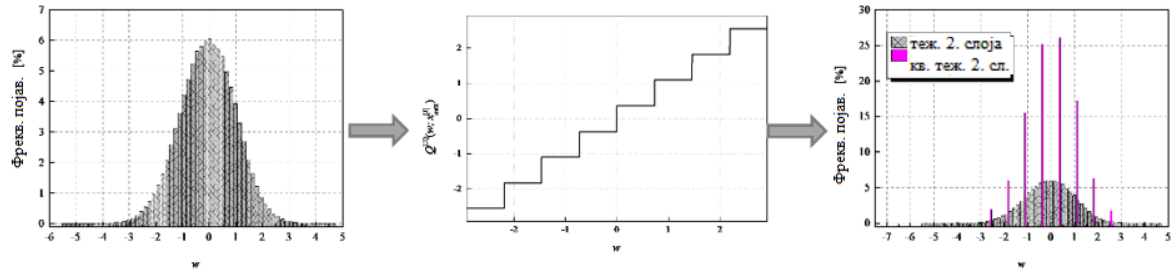
Преносна карактеристика тробитног униформног квантизера за $\Re_g$ избор 3;

Нормализовани хистограм униформно квантованих нормализованих тежина за

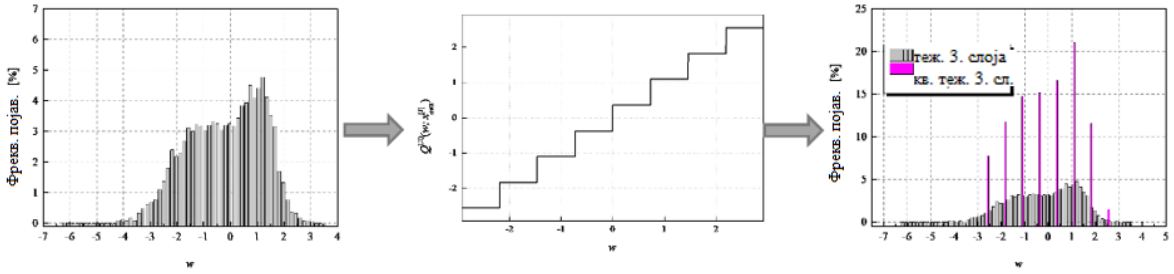
а) слој 1, б) слој 2, ц) слој 3, д) све слојеве MLP.



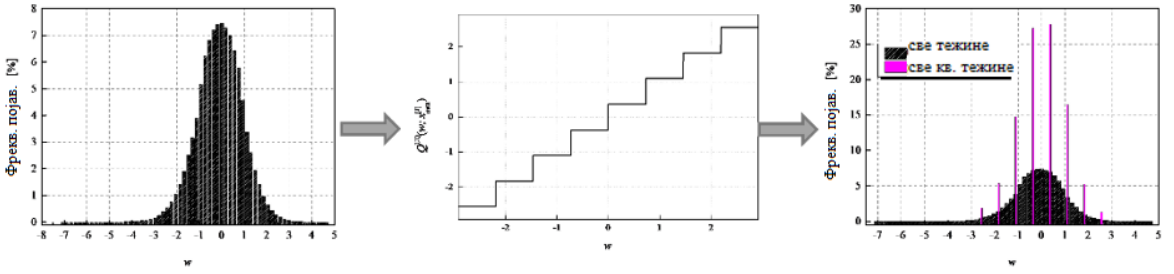
а)



б)



ц)



д)

Слика 4.2.6.4. Нормализовани хистограм за MLP модел трениран на MNIST скупу са нормализованим FP32 тежинама за а) слој 1, б) слој 2, ц) слој 3 д) све слојеве;

Преносна карактеристика тробитног униформног квантизера за $\Re_g$ избор 4;

Нормализовани хистограм униформно квантованих нормализованих тежина за

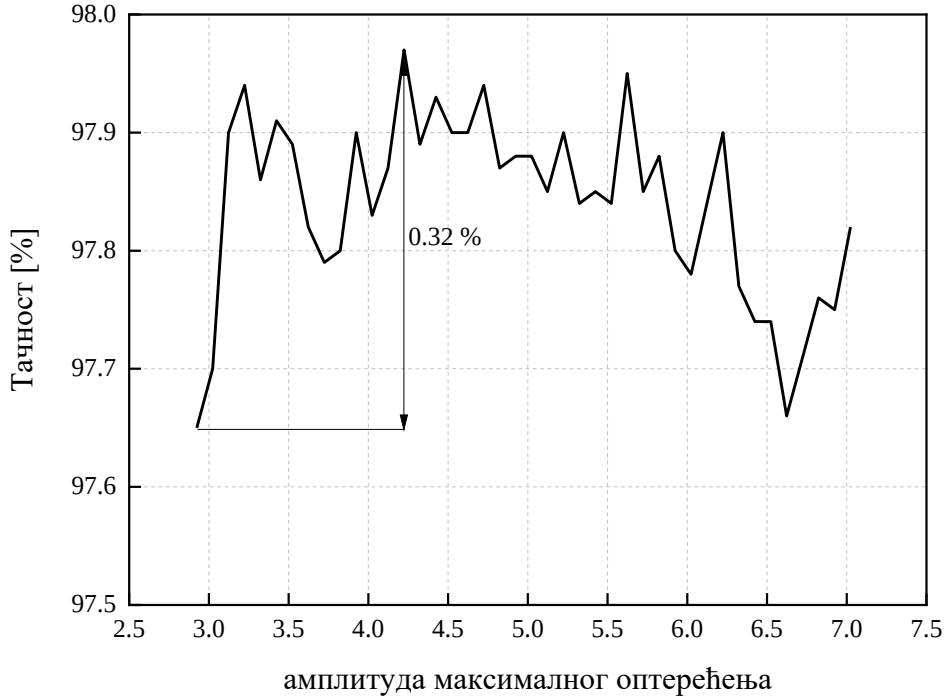
а) слој 1, б) слој 2, ц) слој 3, д) све слојеве MLP.

Посматрајући нормализоване хистограме униформно квантованих тежина за изборе 3 и 4 може се приметити равномернија учесталост коришћења репрезентационих нивоа, што имплицира да користимо пун потенцијал тробитног униформног квантизера. Штавише, ово се може потврдити и упоређивањем обвојница нормализованих хистограма свих нормализованих тежина пре и после квантизације. Такође се може приметити да за изборе 3 и 4, обвојница нормализованог хистограма униформно квантованих нормализованих тежина у великој мери прати обвојницу хистограма свих нормализованих тежина пре квантовања. Ово нам показује да расподела квантованих нормализованих тежина прати расподелу нормализованих тежина пре примене квантизације.

Анализом нормализованих тежина у различитим слојевима NN модела, можемо приметити да слој 3 има највеће одступање од Лапласове расподеле, показујући расподелу која је ближа униформној. Како је униформни квантизер оптималан за униформну расподелу, одговарајућим избором $\Re_g$, предности униформног квантизера за овај слој могу бити значајне.

Нормализовани хистограми нам нуде још један занимљив увид у благу асиметрију нормализованих тежина модела QNN. Ова појава је последица чињенице да је проценат ненегативних тежина након нормализације 50.79 %, што значи да је већи за 1.58 % од процента негативних вредности тежине након нормализације. Само померање расподеле тежина удесно током квантизације, се може анализирати у појединачним и заједничким нормализованим хистограмима нормализованих тежина пре и после квантизације. Асиметрија је највећа на слоју 2, где проценат ненегативних тежина након нормализације износи 51.3649 %, што значи да је већи за 2.7298 % од процента негативних вредности тежина након нормализације.

До сада смо потврдили да у сва четири посматрана избора $\Re_g$, први QNN модел (MLP трениран на MNIST скупу података) постиже стабилне перформансе у погледу тачности, са одступањем од приближно 0.2 %. Да бисмо даље испитали утицај који w_{supp} има на тачност QNN, одредили смо тачност QNN за w_{supp} вредности у опсегу од $x_{max}^{[J]} = 2.9236$ до $|x_{min}| = 7.063787$, са величином корака од 0.1 (видети слику 4.2.6.5.).



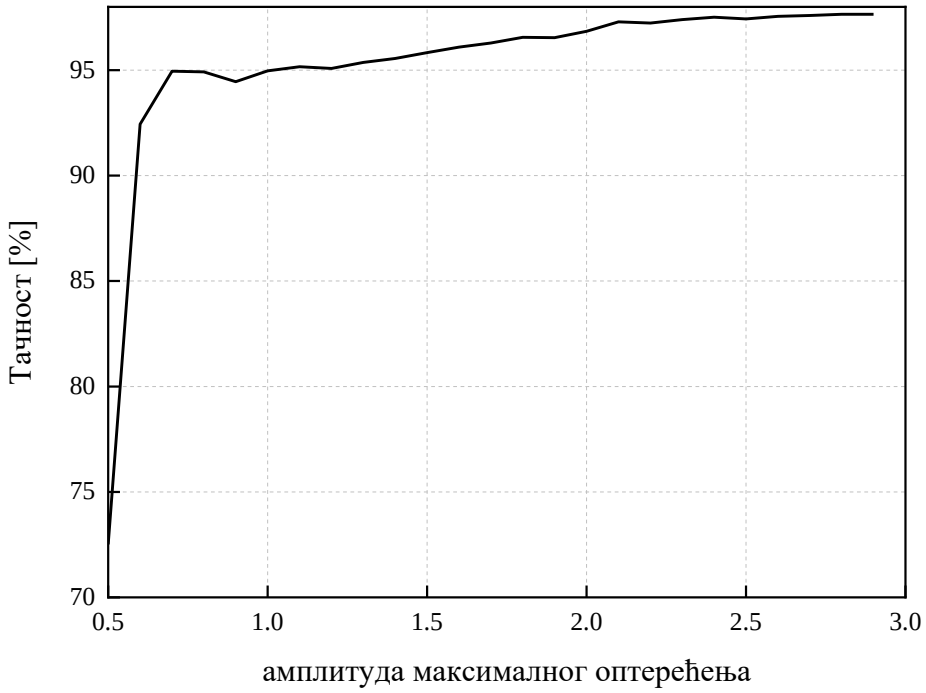
Слика 4.2.6.5. Тачност првог модела QNN (MLP трениран на MNIST скупу) за w_{supp} у интервалу од x_{max}[J] = 2.9236 до $|x_{min}|$ = 7.063787.

Највећа остварена тачност QNN у посматраном опсегу износи 97.97 %, док је најнижа тачност 97.65 %, односно динамика тачности QNN износи 0.32 %. Стога смо још једном потврдили да се за брзину од R = 3 бит/одмерак утицај избора $\Re_g$ на тачност QNN значајно смањује у поређењу са случајем где су идентичне MLP тежине, сачуване у FP32 формату, униформно квантоване брзином од 2 бит/одмерак. Треба напоменути да је динамика тачности приказана на слици 4.2.6.5. одређена у широком, али пажљиво одабраном опсегу посматраних вредности w_{supp}, који смо хеуристички специфицирали у раду [24].

У наставку такође представљамо низ неповољних $\Re_g$ избора за први анализирани NN модел, тј. MLP трениран на MNIST скупу (видети слику 4.2.6.6.). Ако се вредности амплитуде максималног оптерећења крећу од 0.5 до $x_{max}^{[J]}$ = 2.9236 и мењају са

кораком од 0.1 тачност QNN варира од 72.51 % до 97.65 %, што је значајна динамика, те можемо закључити да су сва четири избора приказана у табели 4.2.6.2. пажљиво и погодно одабрана. Поред тога, приметно је да је за вредности веће од $x_{max} = 2.5$ тачност QNN изнад 97 %, због малог утицаја различитих избора $\Re_g$ на тачност QNN, ако је вредност w_{supp} изабрана из одговарајућег интуитивно очекиваног опсега, тј. уколико важи $w_{supp} \geq 2.5$.

Док је тачност NN главна референтна тачка експерименталне анализе, SQNR тробитног униформног квантизера заузима значајно место у нашој теоријској и експерименталној анализи.



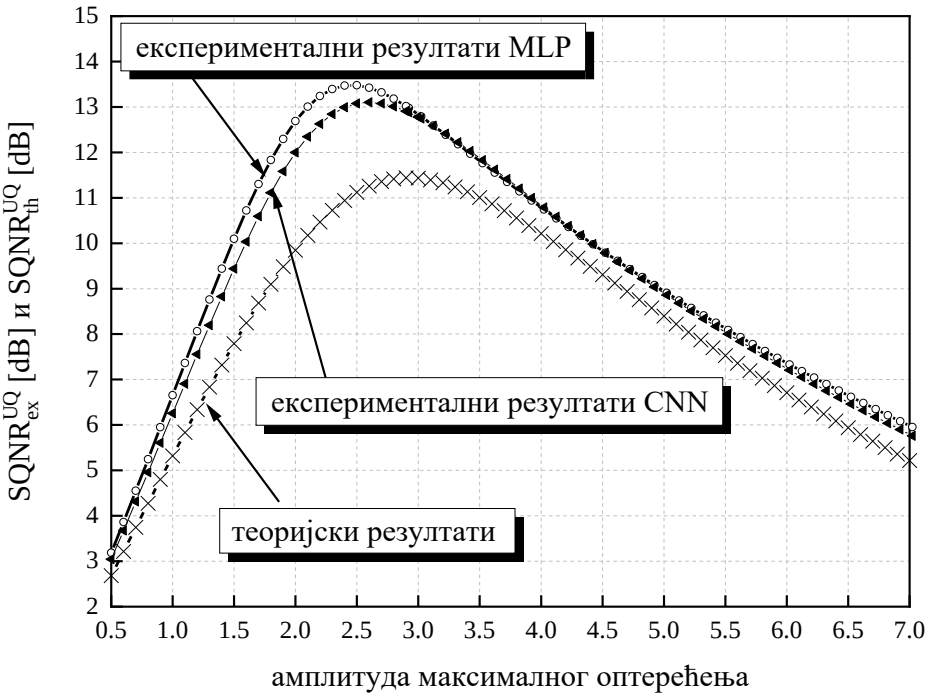
Слика 4.2.6.6. Тачност првог модела QNN (MLP трениран на MNIST скупу) за неповољне изборе амплитуде максималног оптерећења.

Табела 4.2.6.2. је већ показала да за битску брзину од R = 3 бит/одмерак, SQNR вредности имају много већу динамику у поређењу са тачношћу QNN. Да бисмо испитали утицај избора w_{supp} на SQNR у ширем интервалу вредности, одредили смо теоријски и експериментални SQNR тробитног униформног квантизера у опсегу w_{supp} од [0.5, 7.063787]. Поред тога, да бисмо упоредили експериментално одређене вредности за два различита NN модела наведена у раду [24], MLP и CNN, на слици 4.2.6.7. дајемо додатне експерименталне резултате за CNN модел.

Ако поставимо w_{supp} између $x_{max}^{[J]}$ = 2.9236 и $|x_{min}|$ = 7.063787, са величином корака од 0.1, израчунавањем SQNR у 42 тачке динамика експериментално израчунатог SQNR за MLP је 12.9766 dB – 5.9005 dB = 7.0761 dB, што се сматра веома високом вредношћу. За вредности w_{supp} у опсегу између 0.5 и $x_{max}^{[J]}$ = 2.9236 са величином корака од 0.1, динамика SQNR за исти MLP случај је 13.47669 dB – 3.18492 dB = 10.2918 dB.

Очигледно је да избор w_{supp} има велики утицај на остварени SQNR тробитног униформног квантизера. Нешто мања, али и даље значајна динамика експериментално срачунатог SQNR може се уочити за случај модла CNN. Слика 4.2.6.7. приказује разлику у теоријски и експериментално постигнутим SQNR вредностима. На слици 4.2.6.7. се може приметити да теоријска крива SQNR опада по премашивању теоријски оптималне вредности $x_{max}^{[J]}$ = 2.9236. Насупрот томе, експериментално одређена оптимална вредност w_{supp} је мања у односу на теоријску. Ово је последица уже динамике амплитуда реалних нормализованих тежина (видети слику 4.2.6.8.), коришћених у експерименталној анализи, у поређењу са теоријски претпостављеним. Максимум експериментално остварене SQNR вредности у MLP случају износи 13.47669 dB, за амплитуду максималног оптерећења од x_{max} = 2.5, док је експериментално остварена тачност QNN за овај $\Re_g$ избор 97.43 %, што није максимална експериментално остварена тачност QNN. Максимум експериментално остварене SQNR вредности у случају CNN износи 13.10203 dB, за амплитуду максималног оптерећења од x_{max} = 2.6, док експериментално постигнута тачност QNN

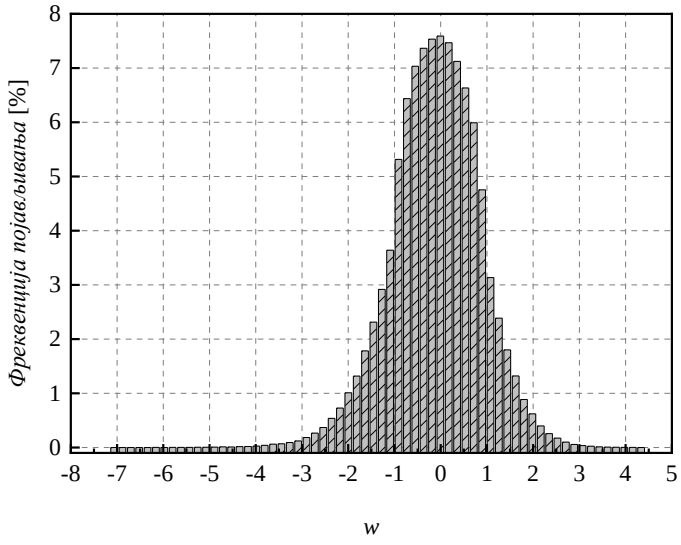
за овај $\Re_g$ избор износи 98.27 %, што такође није максимална експериментално остварена тачност за CNN случај.



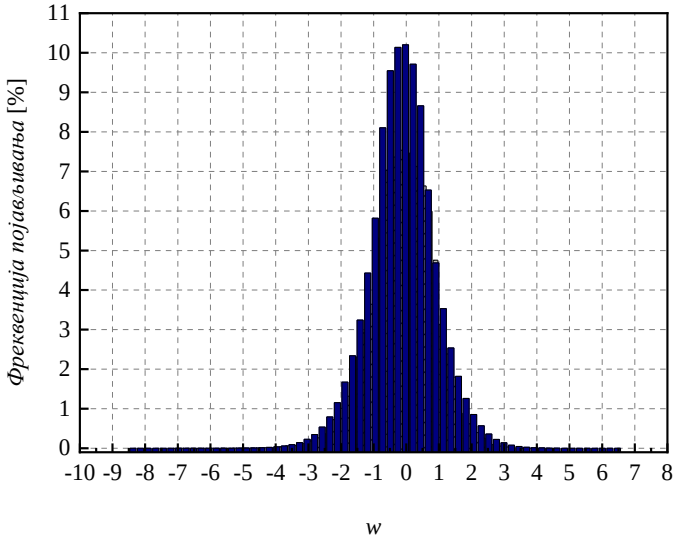
Слика 4.2.6.7. Теоријски и експериментални SQNR за MLP и CNN у широком опсегу w_{supp}.

Може се приметити већа динамика амплитуда тежина за CNN случај у поређењу са MLP случајем (видети слику 4.2.6.8.), тако да је максимум експериментално постигнутих SQNR вредности у CNN случају померен удесно (видети слику 4.2.6.7.). Штавише, може се уочити да се теоријски одређена SQNR крива налази испод обе експериментално одређене SQNR криве. Објашњење лежи у чињеници да се у експерименталној анализи квантују тежине расподеле сличне Лапласовој и из интервала [−7.063787, 4.8371024] за MLP модел, док за CNN модел овај интервал износи [−8.372064, 6.469376], а у теоријској анализи се претпоставља квантизација вредности из бесконачног скупа амплитуда за Лапласов извор, што резултује повећањем дисторзије. Из сличног разлога (шира динамика амплитуда тежина у CNN

случају у поређењу са MLP случајем), експериментално постигнути SQNR је нижи у CNN случају у поређењу са MLP случајем.



а)



б)

Слика 4.2.6.8. Нормализовани хистограм нормализованих тежина за а) MLP, б) CNN (обе мреже су трениране на MNIST сету података).

У поређењу са случајем 1 двобитног униформног квантизера представљеног у одељку 4.2.2 [21], описани тробитни униформни квантизер у случају за поређење (избор 1 за MLP модел) постиже већу тачност на MNIST скупу података валидације за значајних 0.88 %. Ово нас додатно уверава да QNN са униформним квантизером треба ограничити на случајеве избора брзина од 2 или 3 бит/одмерак, јер се очекује да даље повећање битске брзине неће допринети значајном повећању тачности QNN. Другим речима, показали смо да се са посматраним тробитним униформним квантизером може постићи тачност QNN од 97.97 %, што значи да применом описане процедуре пост-тренинг квантизације на MLP модел, деградација тачности од само $98.1\ \% - 97.97\ \% = 0.13\ \%$ се може постићи у поређењу са иницијалним MLP моделом са тежинама сачуваним у FP32 формату.

Уочени случајеви CNN модела и тробитног униформног квантизера су исти као и претходно примењени у MLP анализи. Конкретно, одабрана су четири специфична избора $\mathfrak{R}_g$ и за сваки избор смо одредили SQNR и тачност CNN модела (видети табелу 4.2.6.4.).

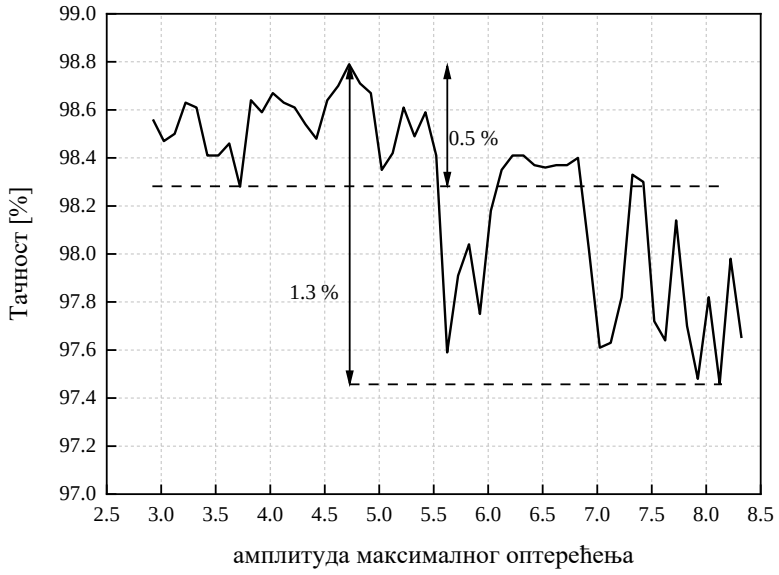
Може се приметити да је тачност QNN за CNN модел релативно стабилна, посебно имајући у виду веома високе вредности $w_{\min}$ и $w_{\max}$ у поређењу са избором 3 и 4.

Табела 4.2.6.4. SQNR и тачност QNN за MLP модел код различитих тробитних униформних квантизера.

$w_{\min} = -8.372064$, $w_{\max} = 6.469376$, $x_{\max}^{[H]} = 2.9408$, $x_{\max}^{[J]} = 2.9236$	**Случај 1**	**Случај 2**	**Случај 3**	**Случај 4**
	$\mathfrak{R}_g$	$\mathfrak{R}_g$	$\mathfrak{R}_g$	$\mathfrak{R}_g$
	$[-w_{\max}, w_{\max}]$	$[w_{\min}, -w_{\min}]$	$[-x_{\max}^{[H]}, x_{\max}^{[H]}]$	$[-x_{\max}^{[J]}, x_{\max}^{[J]}]$
$SQNR_{ex}^{UQ}$[dB]	6.5399	4.0532	12.8594	**12.8812**
$SQNR_{th}^{UQ}$[dB]	5.9846	3.4347	11.4414	**11.4419**
Тачност [%]	98.38	97.64	98.56	**98.56**
Припадност $\mathfrak{R}_g$ [%]	99.9999	100	99.165	99.136

Штавише, референтна тачност (пре квантизације) за CNN и тежине у FP32 формату је очекивано већа у поређењу са MLP, и износи 98.89 % на MNIST скупу валидације. За разлику од тачности QNN, SQNR вредности тробитног униформног квантизера се доста разликују, што је резултат релативно великих апсолутних вредности $w_{\min}$ и $w_{\max}$, које су неколико пута веће од теоријски оптималне вредности амплитуда максималног оптерећења за дату битски брзину и Лапласов извор. Ипак, пошто је флуктоација тачности мања од 1 % (98.56 % – 97.64 % = 0.92 %) за тако различите изборе $\Re_g$, овим се потврђује да су запажања дата за MLP у вези са утицајем $\Re_g$ на перформансе QNN, слична за CNN случај.

Анализом тачности модела QNN представљеног на слици 4.2.6.9. а) за случај CNN, можемо приметити да је динамика тачности QNN, дефинисана као разлика између највеће и најниже постигнуте тачности QNN у веома широком опсегу изабраних амплитуда максималног оптерећења (шири од MLP случаја), износи 1.3 %.



а)

б)

ц)

Слика 4.2.6.9. Тачност QNN након тробитне униформне квантизације за широки интервал вредности за w_{supp} од x_{max}[J] = 2.9236 до $|x_{min}|$ = 8.372064 и за а) CNN модел трениран на MNIST скупу; б) MLP модел трениран на Fashion-MNIST скупу; ц) CNN модел трениран на Fashion-MNIST скупу.

Подсетимо, грубом анализом резултата из табеле 4.2.6.3. за случај имплементације двобитног униформног квантизера, можемо израчунати да је динамика тачности QNN већа и износи 2.39 %. На први поглед, упоређивањем вредности $SQNR_{ex}^{UQ}$ за MLP и $SQNR_{ex}^{UQ}$ за CNN, за сва четири избора (видети табеле 4.2.6.2 и 4.2.6.4), могу се приметити прилично сличне вредности за изборе 3 и 4, $SQNR_{ex}^{UQ}$ (MLP - избори 3 и 4) ≈ $SQNR_{ex}^{UQ}$ (CNN - избори 3 и 4). Међутим, за изборе 1 и 2, очекивано је $SQNR_{ex}^{UQ}$ (MLP избори 1 и 2) > $SQNR_{ex}^{UQ}$ (CNN избори 1 и 2) због веће динамике амплитуда тежина трениране NN за CNN модел у поређењу са оним за MLP модел.

Анализом процента вредности тежина који припадају $\Re_g$ за случајеве CNN и MLP, може се приметити да су ови проценти прилично слични за изборе 1 и 2 (видети табеле 4.2.6.2. и 4.2.6.4.). Припадност $\Re_g$ [%] (MLP избори 1 и 2) ≈ Припадност $\Re_g$ [%] (CNN-избори 1 и 2). Штавише, може се приметити да Припадност $\Re_g$ [%] (MLP - избори 3 и 4) > Припадност $\Re_g$ [%] (CNN - избори 3 и 4). Наиме, проценат вредности тежина унутар $\Re_g$ [%] за MLP и изборе 3 и 4 већи је отприлике за 0.25 %. Обзиром да важи Тачност [%] (MLP - избор 3) ≈ Тачност [%] (MLP - избор 4) и Тачност [%] (CNN - избор 3) ≈ Тачност [%] (CNN - избори 4), може се уочити скоро иста разлика у тачности од 0.9 % у корист CNN за изборе 3 и 4, што је интуитивно и очекивано.

Поред тога, пожељно је и анализирати резултате приказане на слици 4.2.6.9.а) за нешто ужи опсег $\Re_g$. Тачније, ако анализирамо резултате до $w_{\max}$ = 6.469376 ([%] тежина унутар $\Re_g$ за избор 1 је 99.9999%), може се одредити нешто мања динамика тачности QNN.

Из обе анализе спроведене за CNN модел, може се закључити да се прилично уједначена тачност QNN може постићи без обзира на избор $\Re_g$, што није случај са двобитном униформном квантизацијом која је дата за MLP случај у одељку 4.2.2, као и у раду [21] .

Укратко, експериментална анализа је потврдила да за случај R = 3 бит/одмерак, различити избори $\Re_g$ немају једнако велики утицај на тачност QNN као што је то случај

када је битска брзина једнака 2 бит/одмерак. Штавише, показано је у [24] да уколико амплитуда максималног оптерећења узима вредност из правилно дефинисаног широког опсега, за оба наведена NN модела, MLP и CNN, кој су обучени на MNIST скупу података, перформансе QNN су прилично стабилне за различите изборе $\Re_g$, а тачност QNN је у великој мери очувана у поређењу са одговарајућим основним NN моделима.

У наставку су приложени резултати за оба наведена NN модела, MLP и CNN, обучена на Fashion-MNIST скупу података [75]. У табелама 4.2.6.5. и 4.2.6.6. представљени су SQNR и тачност QNN остварени применом различитих тробитних униформних квантизера за MLP и CNN, који су обучени на Fashion-MNIST скупу података, где је $\Re_g$ дефинисан по истим принципима као у претходним табелама.

У поређењу са MLP, CNN модел постиже већу тачност од 98.89 % на MNIST сету за валидацију, као и већу тачност од 91.53 % постигнуту на Fashion-MNIST сету за валидацију, при чему су тежине такође представљене у FP32 формату. Највећа постигнута тачност за посматране $\Re_g$ опсеге у табели 4.2.6.5. износи 88.48 % за избор 4, док је најнижа вредност за избор 2 и износи 87.12 %, односно динамика тачности је 1.36 %. За избор 4 можемо израчунати да деградација тачности услед примене тробитног униформног квантизера износи 88.96 % − 88.48 % = 0.48 %.

Табела 4.2.6.5. SQNR и тачност QNN за MLP модел код различитих тробитних униформних квантизера за Fashion-MNIST.

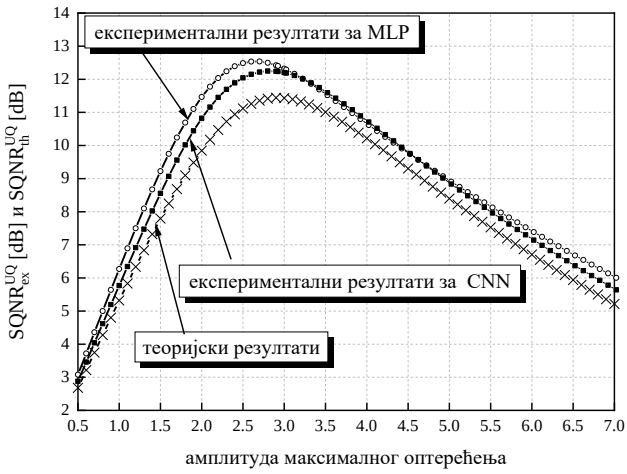
$w_{min} = -9.395458$, $w_{max} = 6.533294$, $x_{max}^{[H]} = 2.9408$, $x_{max}^{[J]} = 2.9236$	**Случај 1**	**Случај 2**	**Случај 3**	**Случај 4**
	$\Re_g$	$\Re_g$	$\Re_g$	$\Re_g$
	$[-w_{max}, w_{max}]$	$[w_{min}, -w_{min}]$	$[-x_{max}^{[H]}, x_{max}^{[H]}]$	$[-x_{max}^{[J]}, x_{max}^{[J]}]$
$SQNR_{ex}^{UQ}$[dB]	6.665	3.1292	12.3843	**12.40141**
$SQNR_{th}^{UQ}$[dB]	5.8894	2.2619	11.4414	**11.4419**
Тачност [%]	87.69	87.12	88.39	**88.48**
Припадност $\Re_g$ [%]	99.997	100	99.027	98.999

Табела 4.2.6.6. SQNR и тачност QNN за CNN модела код различитих тробитних униформних квантизера за Fashion-MNIST.

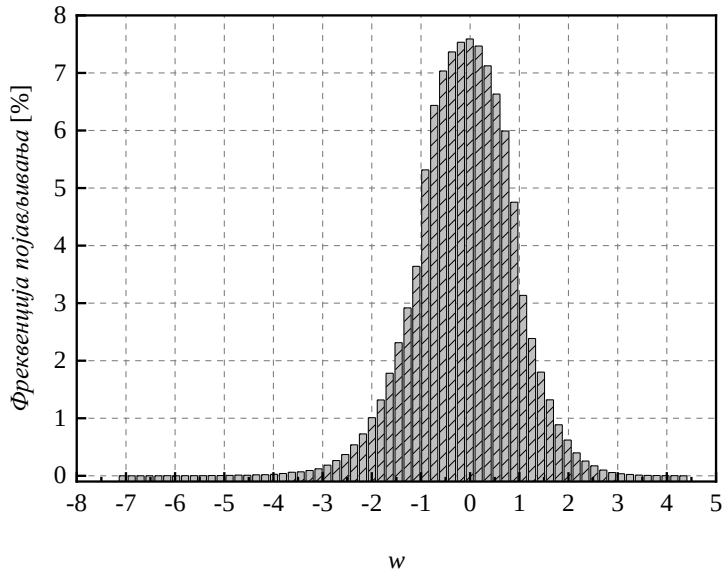
w_{min} = -12.76407909, w_{max} = 7.5561204, $x_{max}^{[H]}$ = 2.9408, $x_{max}^{[J]}$ = 2.9236	**Случај 1** $\Re_g$ $[-w_{max}, w_{max}]$	**Случај 2** $\Re_g$ $[w_{min}, -w_{min}]$	**Случај 3** $\Re_g$ $[-x_{max}^{[H]}, x_{max}^{[H]}]$	**Случај 4** $\Re_g$ $[-x_{max}^{[J]}, x_{max}^{[J]}]$
$SQNR_{ex}^{UQ}$[dB]	4.9069	-0.579776	12.228416	**12.2341143**
$SQNR_{th}^{UQ}$[dB]	4.4615	-0.9315	11.4414	11.4419
Тачност [%]	85.53	84.901	**88.02**	87.97
Припадност $\Re_g$ [%]	99.998	100	98.822	98.784

У табели 4.2.6.6. највећа постигнута тачност износи 88.02 % за избор 3 (деградација услед примене тробитног униформног квантизера износи 91.53 % − 88.02 % = 3.52 %), док је најнижа вредност такође за избор 2, и износи 84.90 %, па је динамика тачности 4.12 %. Избор 2, као што се може приметити из табела 4.2.6.5. и 4.2.6.6. није најпогоднији избор у погледу тачности QNN и SQNR, посебно када је динамика амплитуда нормализованих тежина које се квантују релативно велика. Посматрајући тачност оба QNN модела, (MLP и CNN) за Fashion-MNIST скуп података, (слика 4.2.6.9. б), ц) и за w_{supp} у широком опсегу вредности од $x_{max}^{[J]}$ = 2.9236 до $|x_{min}|$ = 8.372064, динамика тачности QNN износи 2.5 % и 5.5 %, за MLP и CNN, респективно. Поред тога, резултати представљени на слици 4.2.6.9. б) и 4.2.6.9. ц) указују да постоји низ избора за $\Re g$ који могу довести до деградације тачности QNN услед примене тробитног униформног квантизера до 1 % у поређењу са случајем са највећим идентификованом тачношћу QNN. На крају, можемо истаћи да су највеће тачности QNN за посматрани Fashion-MNIST скуп података, а код MLP и CNN модела, постигнуте за w_{supp} = 3.42 и w_{supp} = 3.52, респективно. Ове највеће тачности QNN износе 88.78 % и 89.54 %, за MLP и CNN моделе (за Fashion-MNIST скуп података), респективно. Другим речима, са посматраним тробитним униформним квантизером, и

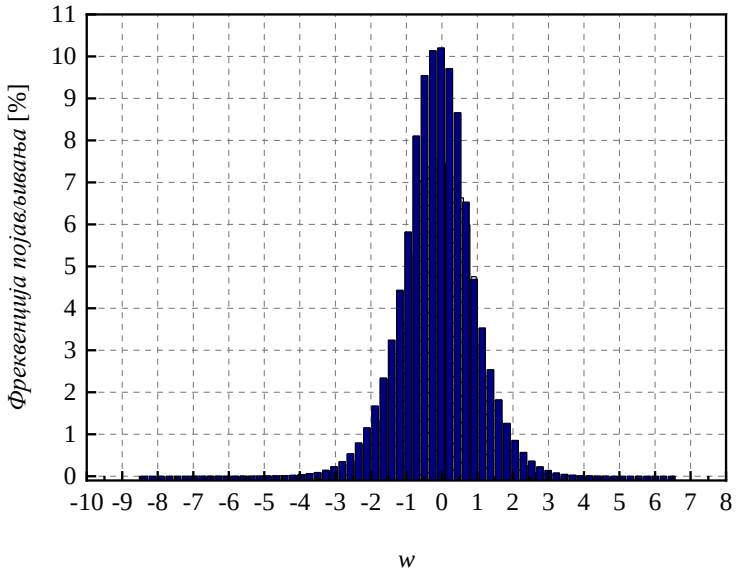
са избором w_{supp} = 3.42 и w_{supp} = 3.52 за одговарајуће случајеве (MLP и CNN), деградација тачности од 88.96 % − 88.78 % = 0.18 % и 91.53 % − 89.54 % ≈ 2 %. Занимљиво је приметити да је тачност за избор 1 смањена за 0.8 % и 2.45 % у поређењу са избором 4, за MLP и CNN, а проценат вредности тежина у $\Re_g$ за избор 1 је већи за 1 % и 1.2 %, респективно. Пошто су експериментално остварени SQNR већи за приближно 2.6dB и 1.6dB за оба модела и избор 1 у корист MNIST сета, постигнуте тачности су веће за приближно 10 % и 13 %. За најужу динамику тежина, односно за избор 4 и за оба NN модела, оба експериментално добијена SQNR су већа за приближно 0.6 dB у корист MNIST сета, док су тачности QNN веће за приближно 9.2 % за MLP и 10.6 % за CNN. За најшири опсег $\Re_g$, тј. за избор 2, и за оба NN модела, експериментално постигнуте SQNR вредности су веће за приближно 2.8 dB и 6.5 dB у корист MNIST, док су тачности QNN веће за приближно 10.7% за MLP и 12.7% за CNN. Са аспекта постигнутог SQNR, очекивано највеће вредности у оба случаја од интереса се постижу за избор 4, док се за избор 2, то јест за непотребно широк опсег $\Re_g$, постижу чак и негативне SQNR вредности, како експерименталне, тако и теоријске.



Слика 4.2.6.10. Теоријске и експерименталне SQNR вредности за MLP и CNN (обе NN трениране на Fashion-MNIST сету) и за широки опсег вредности за w_{supp} тробитног униформног квантизера примењеног за квантизацију у пост-тренинг фази.



а)



б)

Слика 4.2.6.11. Нормализовани хистограм нормализованих FP32 тежина за а) MLP и б) CNN (обе NN трениране на Fashion-MNIST сету).

Као што је већ наглашено, теоријска крива SQNR опада након проласка теоријски оптималне вредности амплитуде максималног оптерећења од 2.9236. Поред тога, као што је претходно закључено за MLP и CNN обучене на MNIST скупу података, у посматраним случајевима са MLP и CNN обученим на Fashion-MNISTскупу података, ова теоријски оптимална вредност w_{supp} се разликује од $w_{supp} = 2.7$ и $w_{supp} = 2.8$, при чему су максимуми експериментално утврђене SQNR вредности од 12.5374 dB и 12.2462 dB постигнути је за MLP и CNN случај, респективно.

Слично објашњењу за MNIST скуп података и MLP и CNN, и овде важи објашњење да су уочене разлике у SQNR вредности последица мање динамике вредности реалних нормализованих тежина (видети слику 4.2.6.11.), коришћених у експерименталној анализи, у поређењу са претпостављеним теоријским вредностима. За $w_{supp} = 2.7$, тачност QNN за MLP обучен на Fashion-MNIST скупу података износи 88.54 %, што није максимална тачност постигнута у експерименту. Сличан закључак се може извести за CNN обучен на Fashion-MNIST скупу података, где смо за $w_{supp} = 2.8$ утврдили да тачност QNN износи 87.73 %.

Као и код MNIST скупа података, у случају Fashion-MNIST скупа података, може се приметити већа динамика амплитуде тежина за CNN случај у поређењу са MLP случајем (види слику 4.2.6.11.). То такође условљава да је максимум експериментално постигнутих SQNR вредности у CNN случају померен удесно на слици 4.2.6.10. Штавише, може се приметити да је теоријски одређена SQNR крива испод обе експериментално одређене SQNR криве, са сличним појашњењима уоченог понашања SQNR крива као за MNIST скуп података.

4.2.7 Анализа деградације тачности неуронске мреже услед униформне квантизације тежина једног или више слојева

За дату амплитуду максималног оптерећења $x_{\max}$, величина квантизационе ћелије двобитног униформног квантзера [23] и правило квантовања су дата са:

$$\Delta = \frac{x_{\max}}{2}, \tag{4.2.7.1}$$

$$Q^{\mathrm{UQ}}(w_j) = \begin{cases} \operatorname{sgn}(w_j)\left(\left\lfloor \frac{|w_j|}{\Delta} \right\rfloor + \frac{1}{2}\right)\Delta, |w_j| \le x_{\max} \\ \operatorname{sgn}(w_j)\frac{3}{2}\Delta, |w_j| > x_{\max} \end{cases}. \tag{4.2.7.2}$$

где w_j означава нормализоване тежине представљене у FP32 формату које треба квантовати коришћењем наведеног правила квантовања, или оставити неквантоване. $Q^{\mathrm{UQ}}\left(w_j\right)$ специфицира излаз двобитног униформног квантизера, за дати улаз w_j. $\lfloor \cdot \rfloor$ означава заокруживање на најближу доњу целу вредност. У нашој анализи, као и у претходним одељцима, тежине представљене у FP32 формату су заправо нормализоване пре квантовања тако да имају нулту средњу вредност и јединичну варијансу, а затим и денормализоване применом инверзног поступка, а након квантовања.

Означимо са $\hat{w}_j$ тежине у фази после тренинга, а пре денормализације, W укупан број нормализованих тежина, а $\{W_i\}_{i=1}^{3}$ број нормализованих тежина на сваком од три слоја наше NN, обучене на MNIST скупу података. У наставку наводимо случајеве од интереса за нашу анализу и изводимо:

$$\textit{Анализа}\ 1: \hat{w}_j = Q^{\mathrm{UQ}}\left(w_j\right),\ j = 1,2,...,W, \tag{4.2.7.3}$$

$$\textit{Анализа}\ 2: \hat{w}_j = \begin{cases} w_j,\ j = 1,2,...,W_1 \\ Q^{\mathrm{UQ}}(w_j),\ j = W_1+1, W_1+2,...,W \end{cases}, \tag{4.2.7.4}$$

$$\textit{Анализа } 3: \hat{w}_j = \begin{cases} w_j, \ j = 1,2,...,W - W_3 \\ Q^{\mathrm{UQ}}(w_j), \ j = W - W_3 + 1, W - W_3 + 2,...,W \end{cases}, \tag{4.2.7.5}$$

$$\textit{Анализа } 4: \hat{w}_j = \begin{cases} Q^{\mathrm{UQ}}(w_j), \ j = 1,2,...,W - W_3 \\ w_j, \ j = W - W_3 + 1, W - W_3 + 2,...,W \end{cases}, \tag{4.2.7.6}$$

где се (4.2.7.3) поклапа са анализом датом у 4.2.2 и овде је дат ради прегледније компарације анализа. У *Анализи* 1, све нормализоване тежине нашег обученог NN су квантоване коришћењем двобитног униформног квнтизера. У *Анализи* 2, нормализоване тежине из првог слоја остају неквантоване, док су нормализоване тежине из другог и трећег слоја квантоване коришћењем двобитног униформног квнтизера. У *Анализи* 3, нормализоване тежине из првог и другог слоја остају неквантоване, док су нормализоване тежине из трећег слоја квантоване коришћењем двобитног униформног квнтизера. На крају, у *Анализи* 4, нормализоване тежине из првог и другог слоја су квантоване коришћењем двобитног униформног квнтизера, док су нормализоване тежине из трећег слоја остављене неквантоване.

Као што је познато, перформансе NN модела зависе од сваког од слојева, док све тежине имају одређену улогу у обезбеђивању исправног или очекиваног излаза модела. Стога, представљамо вишеструке варијације у процесу квантизације након тренинга, да бисмо одредили и истакли и позитивне и негативне утицаје квантизације специфичних слојева на тачност QNN. Поред тога, експерименте ћемо изводити за три различита случаја избора ширине грануларног региона у циљу проналажења најпогодније комбинације параметара униформног квнтизера и слојева који се квантују.

Случај 1 представља избор ширине грануларног региона у којем је $x_{\max} = 4.8371024$, тј. $x_{\max}$ је једнак максималној вредности нормализованих тежина. Сам грануларни регион је симетричан и дефинисан је као $[-x_{\max}, x_{\max}]$. Да бисмо даље прецизирали грануларни регион посматраног униформног квнтизера у случајевима 2 и 3 користимо вредности асимптотски оптималне и оптималне амплитуде максималног оптерећења

за Лапласову функцију густине вероватноће (раније поменути *Hui* (H) и *Jayant* (J) x_{max} [H] = 1.9605 и x_{max} [J] = 2.1748)).

Табела 4.2.7.1. представља део раније приказаних експерименталних резултата у 4.2.2. Поред тачности нашег NN модела обученог на MNIST скупу података без примене квантизације, која износи 98.1 %, дате су и тачности QNN у случају када су све нормализоване тежине квантоване коришћењем двобитног униформног квнтизера за сва три посматрана случаја.

У табели 4.2.7.2. приказани су резултати *Анализе* 2, у којој не примењујемо двобитну униформну квантизацију на први слој нашег NN, док квантујемо нормализоване тежине у слојевима 2 и 3. Остварена тачност је знатно већа у односу на оне приказане у табели 4.2.7.1., што је и очекивано с обзиром да слој 1 садржи око 60 % свих нормализованих тежина (видети табелу 4.2.7.6.). У табели 4.2.7.3. приказани су резултати *Анализе* 3, односно тачност NN модела када двобитни униформни квантизер применимо само на квантовање тежина слоја 3. С обзиром да слој 3 садржи само 0.77 % свих тежина, анализа представљена у табели 3 је очекивани пример лошег избора појединачног слоја за квантизацију. Разлог је што на тај начин, у основи не вршимо скоро никакву компресију на тежинама NN модела, пошто преостала два слоја садрже 99.23 % свих нормализованих тежина.

Табела 4.2.7.1. Тачност QNN модела: све тежине квантоване двобитним униформним квнтизером.

Тачност (FP32) = 98.1 %			
	Случај 1	Случај 2	Случај 3
Тачност [%]	96.97	96.34	96.74

Табела 4.2.7.2. Тачност QNN модела: нормализоване тежине другог и трећег слоја су квантоване.

Тачност (FP32) = 98.1 %			
	Случај 1	Случај 2	Случај 3
Тачност [%]	97.99	97.84	97.84

Табела 4.2.7.3. Тачност NN модела: нормализоване тежине трећег слоја су квантоване

Тачност (FP32) = 98.1 %			
	Случај 1	Случај 2	Случај 3
Тачност [%]	98.01	97.93	97.91

Табела 4.2.7.4. Тачност NN модела: нормализоване тежине првог и другог слоја су квантоване.

Тачност (FP32) = 98.1 %			
	Случај 1	Случај 2	Случај 3
Тачност [%]	97.21	96.74	97.02

Табела 4.2.7.5. Тачност NN модела: Добитак тачности у односу на *Анализу* 1.

Добитак тачности [%]	Случај 1	Случај 2	Случај 3
за *Анализу* 2 у односу на *Анализу* 1	1.02	1.50	1.10
за *Анализу* 3 у односу на *Анализу* 1	1.04	1.59	1.17
за *Анализу* 4 у односу на *Анализу* 1	0.24	0.40	0.28

Табела 4.2.7.6. Број и процентуални удео квантованих тежина по слојевима.

Укупан број тежина NN модела: W = 669 706			
	слој 1	слој 2	слој 3
Број квантованих тежина W_i	401920	262656	5130
Процентуални удео квант. тежина	60.01 %	39.22 %	0.77 %

Табела 4.2.7.4. представља супротну анализу од оне представљене у Табели 4.2.7.3. пошто не примењујемо двобитни униформни квантизер само на слој 3, док се нормализоване тежине из друга два слоја квантују. Ова анализа је посебно важна јер остварујемо повећање тачности QNN модела (0.24 % за случај 1, 0.40 % за случај 2, 0.28 % за случај 3 – видети табелу 4.2.7.5.) у поређењу са анализом представљеном у табели 4.2.7.1. где су све нормализоване тежине квантоване, док се овде квантизација

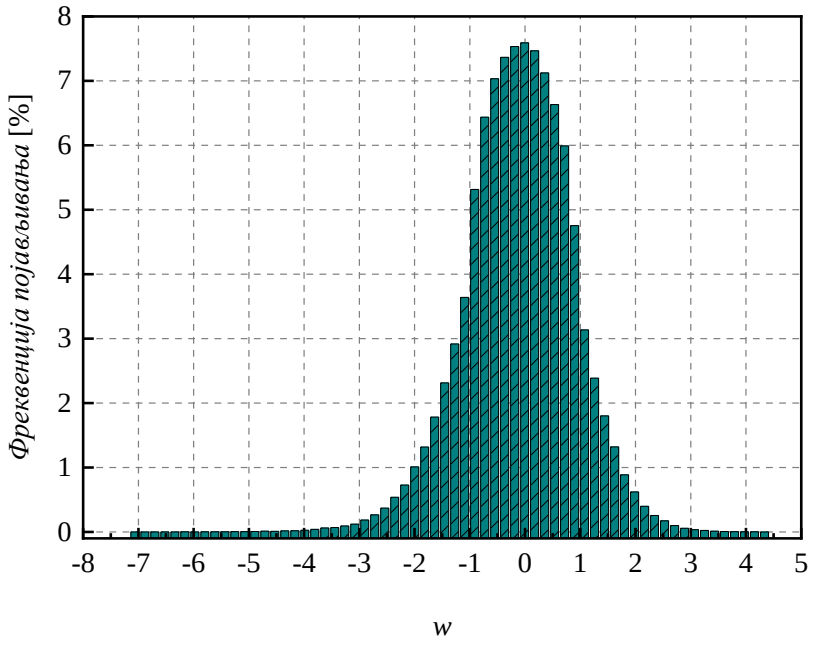
примењује на 99.23 % свих нормализованих тежина. То даље имплицира да би фактор компресије остварен у овој анализи био близак ономе када су све нормализоване тежине квантоване. Просечна битска брзина која се користи за квантизацију тежина у овој анализи може се проценити из:

$$\overline{R}_{|Анализа\ 4} = \frac{2\left(W_1 + W_2\right) + 32\,W_3}{W}. \qquad (4.2.7.7)$$

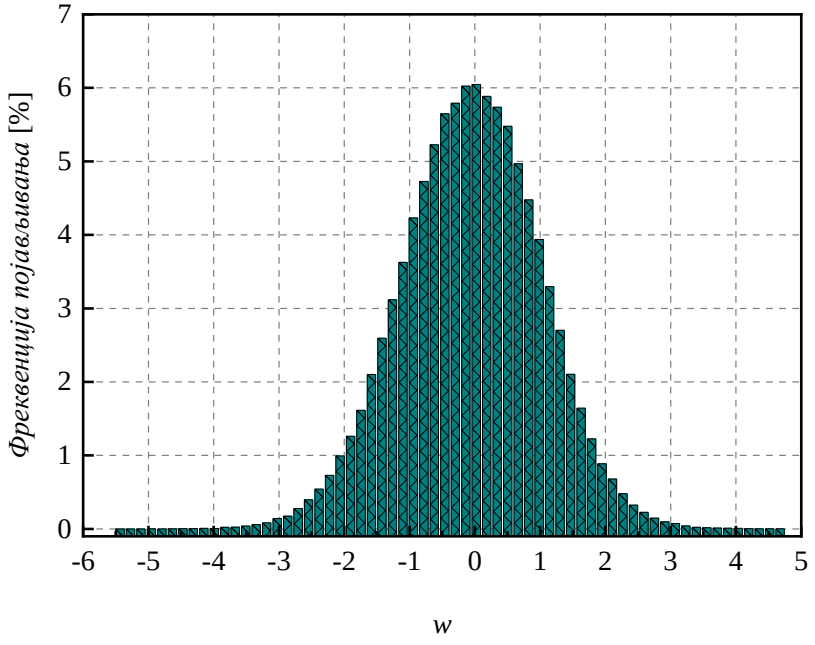
Подсетимо се да W_1, W_2 и W_3 означавају број нормализованих тежина на слојевима 1, 2 и 3, респективно, а W представља укупан број нормализованих тежина. У (4.2.7.7) 2 потиче од брзине која се користи за квантизацију тежина, док је 32-битна брзина која потиче од FP32 формата, те просечна битска брзина модела QNN представљеног у *Анализи* 4 износи приближно 2.23 бит/одмерак. Дакле, можемо закључити да смо повећањем брзине за 0.23 бит/одмерак остварили добитак у тачности модела QNN до 0.40 %. Ово такође указује на важност квантизације примењене на последњем слоју. У наставку ћемо упоредити приказане анализе и елаборирати параметре који утичу на различите изборе слојева за квантизацију.

Поређење у повећању тачности QNN остварено у претходно приказаним анализама приказано је у табели 4.2.7.5. За боље сагледавање резултата треба обратити пажњу на расподелу броја параметара (нормализованих тежина) по слојевима приказану у табели 4.2.7.6. За даље разумевање утицаја квантизације различитих слојева на тачност укупног NN модела, могу се посматрати слике 4.2.7.1. до 4.2.7.3. који представљају нормализоване хистограме нормализованих тежина појединачних NN слојева са тежинама, које се чувају у FP32 формату.

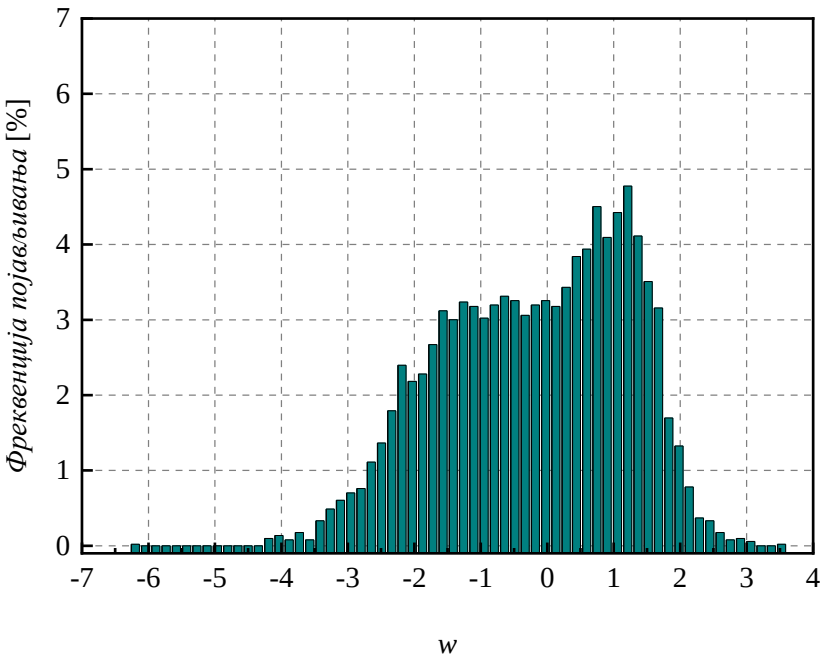
Табела 4.2.7.5. показује да *Анализе* 2 и *Анализе* 3 пружају веома сличан добитак у тачности модела QNN у поређењу са *Анализом* 1, где су све нормализоване тежине квантоване. Узимајући у обзир да у *Анализи* 3 квантујемо само слој 3 са 0.77 % свих тежина, док добијамо веома сличне резултате као у *Анализи* 2, где примењујемо двобитну униформну квантизацију на 39.99 % свих нормализованих тежина из слојева 2 и 3 заједно.



Слика 4.2.7.1. Нормализовани хистограм нормализованих тежина W_1.



Слика 4.2.7.2. Нормализовани хистограм нормализованих тежина W_2.



Слика 4.2.7.3. Нормализовани хистограм нормализованих тежина W_3.

Приметно је да је избор параметара који ће се квантовати у *Анализи* 3 много кориснији у поређењу са онима који се користе у *Анализи* 2. Другим речима, квантовани слој 3 уводи веома сличну деградацију тачности нашег модела QNN као у случају квантовања оба слоја 2 и 3 (видети нормализовани хистограм нормализованих тежина слоја 3 представљен на сл. 4.2.7.3.). Очигледно је да овај хистограм има значајно одступање од Лапласове функције густине вероватноће која се користи за моделовање улазног сигнала приликом пројектовања и оптимизације x_{max} нашег двобитног унифромног квантизера. Према томе, расподела нормализованих тежина у слоју 3, не иде у корист нашој спецификацији грануларног региона квантизера и применом квантизације на слоју 3 убацујемо много већу деградацију у тачности модела QNN по броју тежина, у поређењу са слојевима 1 и 2, чији су нормализовани хистограми значајно ближи Лапласовој функцији густине вероватноће. Можемо закључити да број параметара по слоју не игра пресудну улогу у деградацији тачности модела QNN, док расподеле вредности нормализованих тежина по слојевима имају

велики утицај на успешну примену нискобитне квантизације. До сличног закључка се долази упоређивањем тачности модела QNN у *Анализама* 1 и 4. Може се приметити да непримењивањем квантизације на слој 3 добијамо повећање у тачности QNN до 0.40 % у поређењу са *Анализом* 1, док *Анализа* 4 нуди најефикаснију употребу битске брзине и постиже значајан добитак у тачности, а врши се квантизација 99.23 % свих нормализованих тежина.

У овом одељку смо, као и у раду [23], анализирали различите сценарије у процесу квантизације тежина QNN у фази након обуке, да бисмо одредили и истакли и позитивне и негативне утицаје нискобитне униформне квантизације специфичних слојева на тачност QNN. Указали смо на чињеницу да се по идентификацији најосетљивијег слоја NN може додатно параметризовати двобитни униформни двобитни квантизер за уочени слој на прецизнији начин, доприносећи мањој деградацији тачности QNN. Утврђивање приоритета примене квантизације слоја NN, довело нас је до закључка, да је најважнији слој за квантизацију први слој NN, са највећим бројем тежина, док је последњи слој, иако садржи најмањи број тежина, други по важности слој за примену квантизације тежина. Овај закључак би могао бити посебно користан за прелиминарно примену квантизације у већим NN архитектурама уз примену квантизације мешовите прецизности, тј. битске брзине.

5. Пројектовање нискобитних неуниформних квантизера и њихова примена код неуронских мрежа

5.1.1 Нови модели двобитних неуниформних квантизера за компресију тежина неуронских мрежа

У овом поглављу описујемо два нова модела неуниформног квантизера који домишљато користе једно од два својства најједноставнијег модела квантизера, тј. униформног квантизера. Оба неуниформна квантизера имају исто правило квантизације за одређивање ширине грануларног региона, али имају различите почетне поставке у смислу ширине ћелије у поређењу са униформним квантизером. Први квантизер, једноставни квантизер степена два, који смо именовали - SPTQ (*Simple Power of Two Quantizer*), дефинише ширине ћелија које се множе степеном броја два, док су репрезентациони нивои SPTQ, као и код униформног квантизера, центрирани на средини квантизационих ћелија. Други квантизер, модификовани SPTQ, тј. MSPTQ (*Modified Simple Power of Two Quantizer*), је побољшан и перформансно конкурентнији модел квантизера, код кога су нивои одлуке квантизера центрирани између најближих репрезентационих нивоа, слично као код униформног квантизера. Ова својства чине дизајн нових неуниформних квантизера релативно једноставним. За разлику од униформних квантизера, квантизационе ћелије MSPTQ нису једнаке ширине и репрезентациони нивои нису центрирани на средини квантизационих ћелија. Прецизније, у SPTQ моделу, квантизациона ћелија најближа средњој вредности раподеле нормализованих тежина, тј. нули, има ширину Δ, док је ширина следеће ћелије, која лежи у остатку грануларног интервала, 2Δ, тако да је за SPTQ грануларни интервал дефинисан као $[-3\Delta, 3\Delta]$. Оба горе поменута модела квантизације, SPTQ и MSPTQ, прате исто правило за дефинисање ширине грануларног региона $[-3\Delta, 3\Delta]$ или $[-3\Delta^{mod}, 3\Delta^{mod}]$, респективно, за $R = 2$ бит/одмерак, док се разликују у начину специфицирања нивоа одлуке и репрезентационих нивоа.

Детаљније, у овом поглављу најпре дајемо мотивацију спроведене анализе, затим описујемо процедуру пројектовања SPTQ и MSPTQ и вршимо њихову оптимизацију за претпостављену Лапласову расподелу тежина. Након тога, вршимо квантизацију нормализованих тежина у пост-тренинг фази применом SPTQ и MSPTQ, проучавамо одрживост тачности QNN и указујемо на предности имплементације ова два модела неуниформних квантизера у односу на случај када се за исти класификациони задатак у QNN користи униформни квантизер са једнаким бројем квантизационих ћелија. Наиме, не само да описујемо два нова модела неуниформних квантизера, већ и вршимо њихову оптимизацију тако да дисторзија буде минимална, односно да је SQNR максималан. Затим примењујемо ове неуниформне квантизере у квантизацији тежина нашег NN модела и испитујемо очување тачности за исти задатак класификације као што је наведено у 4. поглављу [21], [24]. Да бисмо обезбедили коректно поређење са резултатима из 4. поглавља [21], [24], у овом поглављу се претпостављају не само идентична MLP архитектура, већ и идентичне тежине (чуване у FP32 формату) које су неуниформно квантоване коришћењем само два бита по одмерку и према потпуно новим правилима неуниформне квантизације. Оба описана модела неуниформних квантизера могу бити посебно значајна код уређаја са ограниченом меморијом, где су једноставна решења прихватљиве тачности од пресудне важности.

Главна мотивација за проналажење погодног двобитног неуниформног квантизера произилази из чињенице да су неуниформни квантизери погоднији за неуниформне расподеле, као што је Лапласова расподела. Конкретно, као у [21], [24], претпостављамо расподелу сличну Лапласовој за експерименталну расподелу тежина и Лапласову функцију густине вероватноће за теоријску расподелу тежина. Подсетимо на закључке о двобитним униформним квантизерима и потврђеној премиси за двобитни униформни квантизер, датим у раду [21], да се параметризација квантизера показала кључном не само за перформансе самог квантизера, већ и за тачност QNN због само 4 расположива репрезентациона нивоа. Побољшање перформанси у односу на униформни квантизер се релативно лако постиже помоћу неуниформне квантизације са већим бројем битова, што није случај са нискобитном неуниформном

квантизацијом због малог броја расположивих репрезентационих нивоа. У наставку најпре дајемо детаљан опис SPTQ и MSPTQ модела квантизера за претпостављену Лапласову функцију густине вероватноће.

5.1.2 Дизајн симетричног SPTQ за Лапласов извор

Подсетимо да је квантизација свеприсутна у обради сигнала и да представља мапирање континуалних вредности сигнала у дискретни скуп вредности од N квантизационих или репрезентационих нивоа [2]. Примарни циљ квантизације је да се минимизира дисторзија, односно одступање квантованог сигнала ($Q_N(X)$) у од оригиналног (X), за дато N и брзину протока R, где је $R = \log_2 N$ [2]

$$D = E\left[\left(X - Q_N\left(X\right)\right)^2\right]. \tag{5.1.2.1}$$

Поступком неуниформне квантизације, као и код униформне квантизације, опсег амплитуде улазног сигнала се дели на грануларни регион $\Re_g$ и регион прекорачења $\Re_o$ (погледати слику 5.1.2.1. за SPTQ). За било који симетрични квантизер, ови региони су раздвојени амплитудама максималног оптерећења које су означене са $-x_{\max}$ и $x_{\max}$, респективно [2]. Грануларни регион $\Re_g$ је дефинисан као:

$$\Re_g = \bigcup_{i=-N/2}^{-1} \Re_i \cup \bigcup_{i=1}^{N/2} \Re_i = \left[-x_{\max}, x_{\max}\right] , \tag{5.1.2.2}$$

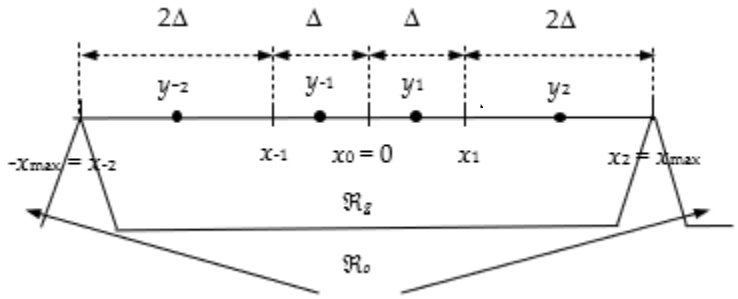
и састоји се од N непреклапајућих квантизационих ћелија ограничених ширина, где је i-та ћелија дефинисана као:

$$\Re_i = \{x \mid x \in [-x_{\max}, x_{\max}], Q_N\left(x\right) = y_i\}, \Re_i \cap \Re_j = \varnothing,\ i \neq j\,. \tag{5.1.2.3}$$

y_i означава i-ти репрезентациони ниво а $\left\{\Re_i\right\}_{i=-N/2}^{-1}$ и $\left\{\Re_i\right\}_{i=1}^{N/2}$ означавају квантизационе ћелије из региона негативних и позитивних вредности амплитуда, које су симетрично постављене око нуле, тј. око средње вредности. Симетрија важи и за ћелије прекорачења, односно за пар квантизационих ћелија неограничене ширине у региону прекорачења $\Re_o$, који је дефинисан као:

$$\Re_0 = \{x \mid x \notin [-x_{\max}, x_{\max}], Q_N(x) = y_{N/2}, x > 0 \vee Q_N(x) = y_{-N/2}, x < 0\}. \qquad (5.1.2.4)$$

Квантизационе ћелије двобитног SPTQ, чији дизајн овде разматрамо, су неједнаке ширине, те је квантизер неуниформни [2].



Слика 5.1.2.1. Грануларни регион $\Re_g$ и регион прекорачења $\Re_o$ симетричног двобитног SPTQ.

Означимо са Δ ширину ћелије нашег симетричног двобитног SPTQ која је најближа средњој вредности функције густине вероватноће, док је ширина њој суседне ћелије двоструко већа (видити слику 5.1.2.1, само позитивну половину амплитудног региона).

За двобитни SPTQ важи:

$$x_{\max} = \Delta + 2\Delta, \qquad (5.1.2.5)$$

$$\Delta = x_{\max} / 3. \qquad (5.1.2.6)$$

Нивои одлучивања нашег двобитног SPTQ одређени су као:

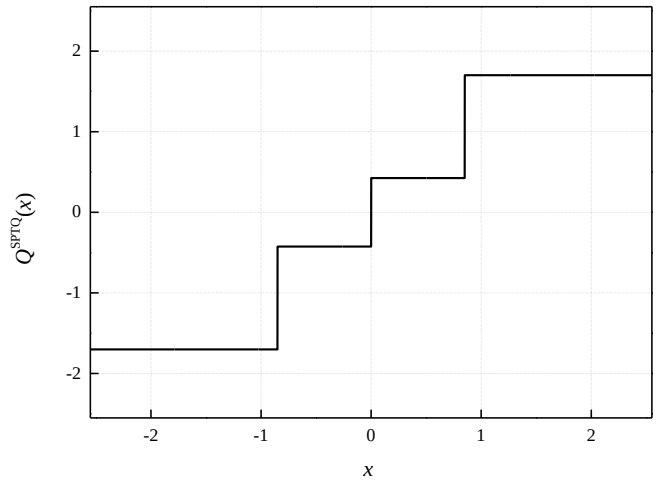
$$x_i = \left(2^i - 1\right)\Delta,\ x_{-i} = -x_i,\ i \in \{0, 1, 2\}, \qquad (5.1.2.7)$$

Кодна књига нашег двобитног SPTQ, $Y^{\text{SPTQ}} \equiv \{y_{-2}, y_{-1}, y_1, y_2\} \subset \mathbb{R}$ садржи $N = 4$ репрезентациона нивоа y_i (погледати слику 5.1.2.1.), који су позиционирани на средини квантизационих ћелија:

$$y_i = \frac{(x_{i-1} + x_i)}{2} = (2^{i-1} + 2^{i-2} - 1)\Delta,\ y_{-i} = -y_i =,\ i \in \{1, 2\}. \qquad (5.1.2.8)$$

Из израза (5.1.2.5) - (5.1.2.8) може се закључити да $x_{max} = x_{max}^{SPTQ}$ или величина корака Δ, у потпуности одређују нивое одлуке x_i и репрезентационе нивое y_i описаног двобитног SPTQ.

Преносна карактеристика симетричног двобитног SPTQ, $Q^{SPTQ}(x; x_{max})$ на слици 5.1.2.2. је представљена за x_{max} = 2.5512, где је нотација [J], као и у ранијим поглављима, потиче од имена аутора [2])



Слика 5.1.2.2. Преносна карактеристика SPTQ $Q^{SPTQ}(x)$ за $[-x_{max}^{SPTQ}, x_{max}^{SPTQ}]$ = [-2.5512, 2.5512].

Полазимо од основне дефиниције дисторзије дате једначином (1) у [2], где су грануларна дисторзија и дисторзија прекорачења за симетрични двобитни SPTQ дате као:

$$D_g^{SPTQ} = 2\sum_{i=1}^{2}\int_{x_{i-1}}^{x_i}(x-y_i)^2\,p(x)dx, \qquad (5.1.2.9)$$

$$D_o^{SPTQ} = 2\int_{x_2}^{\infty}(x-y_2)^2\,p(x)dx, \qquad (5.1.2.10)$$

$$D^{\mathrm{SPTQ}} = 2\sum_{i=1}^{3}\int_{x_{i-1}}^{x_i} x^2 p(x)dx - 4\left(\sum_{i=1}^{2} y_i \int_{x_{i-1}}^{x_i} xp(x)dx + y_2 \int_{x_2}^{\infty} xp(x)dx\right) + 2\left(\sum_{i=1}^{2} {y_i}^2 \int_{x_{i-1}}^{x_i} p(x)dx + {y_2}^2 \int_{x_2}^{\infty} p(x)dx\right). \quad (5.1.2.11)$$

Како би поједноставили наше извођење означавамо горњу границу интеграла у једначини (5.1.2.11) са $x_3 = \infty$. Укупну дисторзију нашег симетричног двобитног SPTQ за Лапласову функцију густине вероватноће налазимо као:

$$D^{\mathrm{SPTQ}} = 1 - \sqrt{2}\left[y_1 + \left(\sqrt{2}x_1 + 1\right)\exp\left\{-\sqrt{2}x_1\right\}\left(y_2 - y_1\right)\right] + y_1^2 + \exp\left\{-\sqrt{2}x_1\right\}\left(y_2^2 - y_1^2\right), \quad (5.1.2.12)$$

$$D^{\mathrm{SPTQ}} = 1 + {y_1}^2 - \sqrt{2}y_1 + \left[\left(y_2 - y_1\right)\left(y_2 + y_1 - 2x_1 - \sqrt{2}\right)\right]\exp\left\{-\sqrt{2}x_1\right\}. \quad (5.1.2.13)$$

Даља замена (5.1.2.7) и (5.1.2.8) у (5.1.2.13) даје:

$$D^{\mathrm{SPTQ}} = 1 - \frac{\sqrt{2}}{2}\Delta + \frac{\Delta^2}{4} + \left[\frac{3}{4}\Delta^2 - \frac{3\sqrt{2}}{2}\Delta\right]\exp\left\{-\sqrt{2}\Delta\right\}. \quad (5.1.2.14)$$

Минимизирањем дисторзије по Δ, односно изједначавањем првог извода тако добијене дисторзије (5.1.2.14) по Δ са нулом налазимо:

$$\frac{\partial D^{\mathrm{SPTQ}}}{\partial \Delta} = 0, \quad (5.1.2.15)$$

односно

$$\Delta - \sqrt{2} + \left(\frac{3\sqrt{2}}{2}\Delta^2 - 9\Delta + 3\sqrt{2}\right)\exp\left\{-\sqrt{2}\Delta\right\} = 0. \quad (5.1.2.16)$$

Даље Δ налазимо итеративно из:

$$\Delta^{(i+1)} = \sqrt{2} + \left(\frac{3\sqrt{2}}{2}\left(\Delta^{(i)}\right)^2 - 9\Delta^{(i)} + 3\sqrt{2}\right)\exp\left\{-\sqrt{2}\Delta^{(i)}\right\}. \quad (5.1.2.17)$$

док је други извод дисторзије по Δ:

$$\frac{\partial^2 D^{\mathrm{SPTQ}}}{\partial \Delta^2} = \frac{1}{2}\left[1 + 3\exp\left\{-\sqrt{2}\Delta\right\}\left(\left(\Delta - 2\sqrt{2}\right)^2 - 3\right)\right]. \qquad (5.1.2.18)$$

Тривијално је закључити да за $\Delta \le 2\sqrt{2} - \sqrt{3}$ и $\Delta \ge 2\sqrt{2} + \sqrt{3}$ важи $\partial^2 D^{\mathrm{SPTQ}} / \partial \Delta^2 \ge 1/2$. Посебну пажњу треба обратити на вредности за Δ из интервала $2\sqrt{2} - \sqrt{3} < \Delta < 2\sqrt{2} + \sqrt{3}$, где можемо очекивати минимум $\partial^2 D^{\mathrm{SPTQ}} / \partial \Delta^2$. Даљим тражењем извода израза (5.1.2.18) по Δ и изједначавањем резултата са нулом

$$\frac{\partial}{\partial \Delta}\left[\frac{1}{2}\left[1 + 3\exp\left\{-\sqrt{2}\Delta\right\}\left(\left(\Delta - 2\sqrt{2}\right)^2 - 3\right)\right]\right] = 0, \qquad (5.1.2.19)$$

налазимо:

$$\Delta^2 - 5\sqrt{2}\Delta + 9 = 0, \qquad (5.1.2.20)$$

и одређујемо корене једначине (5.1.2.20) као $\Delta_{1,2} = \left(5 \pm \sqrt{7}\right)/\sqrt{2}$.

Како $\Delta_2 = \left(5 + \sqrt{7}\right)/\sqrt{2}$ не припада интервалу Δ вредности $2\sqrt{2} - \sqrt{3} < \Delta < 2\sqrt{2} + \sqrt{3}$, минимална вредност израза (5.1.2.20) се постиже за $\Delta_1 = \left(5 - \sqrt{7}\right)/\sqrt{2}$ и износи 0.263. Дакле, можемо закључити да је D^{SPTQ} конвексна функција по Δ. Штавише, да бисмо потврдили да је резултат итеративног поступка јединствена оптимална вредност за Δ, у одељку са нумеричким резултатима (5.1.5), детаљно анализирамо резултате нумеричке оптимизације дисторзије по Δ.

5.1.3 Дизајн симетричног MSPTQ за Лапласов извор

Претпоставимо исто правило квантизације за дефинисање грануларног региона MSPTQ, тј. $[-3\Delta^{\mathrm{mod}}, 3\Delta^{\mathrm{mod}}]$, где је Δ^{mod} ширина ћелије која је најближа средњој вредности Лапласове функције густине вероватноће, као и исто правило за дефинисање репрезентационих нивоа, овде означених са и y_1^{mod} и y_2^{mod}

$$y_i^{\mathrm{mod}} = (2^{i-1} + 2^{i-2} - 1)\Delta^{\mathrm{mod}}, \; y_{-i}^{\mathrm{mod}} = -y_i^{\mathrm{mod}} =, \; i \in \{1, 2\}, \qquad (5.1.3.1)$$

где *mod* означава модификацију. Претпоставимо даље додатну спецификацију дизајна MSPTQ која подразумева да је ниво одлучивања квантизера центриран између најближих репрезентационих нивои (слично као код униформног квантизера):

$$x_1^{\text{mod}} = \frac{y_1^{\text{mod}} + y_2^{\text{mod}}}{2}. \tag{5.1.3.2}$$

Повећање ширине прве квантизационе ћелије MSPTQ, која је најближа средњој вредности функције густине вероватноће, узрокује сужавање суседне квантизационе ћелије (видети слику 5.1.3.1.), а у складу са заједничким правилом квантизације за ширину грануларног региона за оба квантизера $[-3\Delta, 3\Delta]$ или $[-3\Delta^{\text{mod}}, 3\Delta^{\text{mod}}]$. Услов наведен у изразу (5.1.3.2), један је од предуслова за дизајн MSPTQ, пошто је потребно одредити и величину Δ^{mod}, и то из услова минималне дисторзије по Δ^{mod}. За дати $x_{\max}{}^{\text{MSPTQ}} = 3\Delta^{\text{mod}} = x_2{}^{\text{mod}}$ а у складу са условима (5.1.3.1) и (5.1.3.2), можемо одредити кодну књигу $Y^{\text{MSPTQ}} \equiv \left\{ y_{-2}^{\text{mod}}, y_{-1}^{\text{mod}}, y_1^{\text{mod}}, y_2^{\text{mod}} \right\} \subset \mathbb{R}$ нашег двобитног MSPTQ и ниво одлучивања $x_1{}^{\text{mod}}$.

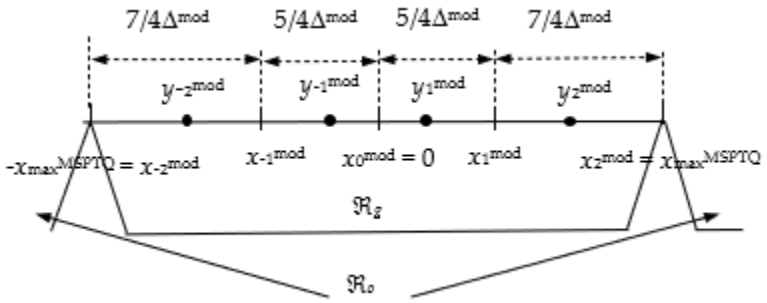
Да бисмо јасно разликовали два модела неуниформних квантизера, које предложемо у овом поглављу, у табели 5.1.3.1. сумирани су главни параметри који једнозначно и недвосмислено описују наше неуниформне квантизере: SPTQ (Табела 5.1.3.1 а)) и MSPTQ (Табела 5.1.3.1 б)). Подсетимо, репрезентациони нивои SPTQ и MSPTQ прате исто правило $y_1 = \Delta/2$, $y_1{}^{\text{mod}} = \Delta^{\text{mod}}/2$ и $y_2 = 2\Delta$, $y_2{}^{\text{mod}} = 2\Delta^{\text{mod}}$, при чему је главна разлика у специфицирању нивоа одлучивања x_1 и $x_1{}^{\text{mod}}$.

За одређивање дисторзије MSPTQ, применом (5.1.2.13) и услова специфицираног у (5.1.3.2), налазимо:

$$D^{\text{MSPTQ}} = 1 + \left(y_1^{\text{mod}} \right)^2 - \sqrt{2} y_1^{\text{mod}} - \left[\sqrt{2} \left(y_2^{\text{mod}} - y_1^{\text{mod}} \right) \right] \exp\left\{ -\sqrt{2} x_1^{\text{mod}} \right\}, \tag{5.1.3.3}$$

док се зависност дисторзије MSPTQ само од Δ^{mod}, а не од репрезентационог нивоа и нивоа одлуке, можемо приказати као:

$$D^{\mathrm{MSPTQ}} = 1 + \frac{\left(\Delta^{\mathrm{mod}}\right)^2}{4} - \frac{\sqrt{2}}{2}\Delta^{\mathrm{mod}}\left[1 + 3\exp\left\{-\frac{5\sqrt{2}\Delta^{\mathrm{mod}}}{4}\right\}\right]. \qquad (5.1.3.4)$$



Слика 5.1.3.1. Грануларни регион $\Re_g$ и регион прекорачења $\Re_o$ симетричног двобитног MSPTQ.

Табела 5.1.3.1. Нивои одлучивања и репрезентациони нивои двобитног а) SPTQ, б) MSPTQ.

Квантизер	x_0	x_1	$x_2=x_{\max}^{\mathrm{SPTQ}}$	y_1	y_2
SPTQ	0	Δ	3Δ	$1/2\Delta$	2Δ

а)

Квантизер	x_0^{mod}	x_1^{mod}	$x_2^{\mathrm{mod}}= x_{\max}^{\mathrm{MSPTQ}}$	y_1^{mod}	y_2^{mod}
MSPTQ	0	$5/4\ \Delta^{\mathrm{mod}}$	$3\Delta^{\mathrm{mod}}$	$1/2\Delta^{\mathrm{mod}}$	$2\Delta^{\mathrm{mod}}$

б)

Изједначавањем првог извода дисторзије, дате изразом (5.1.3.4), по Δ^{mod} са нулом:

$$\frac{\partial D^{\mathrm{MSPTQ}}}{\partial \Delta^{\mathrm{mod}}} = 0, \qquad (5.1.3.5)$$

налазимо:

$$\frac{\Delta^{\mathrm{mod}}}{\sqrt{2}}\left(15 + 2\exp\left\{\frac{5\sqrt{2}\Delta^{\mathrm{mod}}}{4}\right\}\right) - \left(6 - 2\exp\left\{\frac{5\sqrt{2}\Delta^{\mathrm{mod}}}{4}\right\}\right) = 0, \qquad (5.1.3.6)$$

те се Δ^{mod} може итеративно одредити као:

$$\Delta^{\mathrm{mod}^{(i+1)}} = \sqrt{2}\left(1 - \frac{9}{15 + 2\exp\left\{\frac{5\sqrt{2}\Delta^{\mathrm{mod}^{(i)}}}{4}\right\}}\right). \qquad (5.1.3.7)$$

Одређивањем Δ^{mod} можемо израчунати укупу дисторзију MSPTQ из израза (5.1.3.4). Затим можемо наћи други извод тако добијене дисторзије по Δ^{mod}:

$$\frac{\partial^2 D^{\text{MSPTQ}}}{\partial\left(\Delta^{\text{mod}}\right)^2} = \frac{1}{2}\left[1+\left(15-\frac{75\sqrt{2}}{8}\right)\exp\left\{-\frac{5\sqrt{2}\Delta^{\text{mod}}}{4}\right\}\right]. \tag{5.1.3.8}$$

Како важи $15-75\sqrt{2}/8>0$ и стога важи $\partial^2 D^{\text{MSPTQ}}/\partial\left(\Delta^{\text{mod}}\right)^2 \geq 1/2$, можемо закључити да је D^{MSPTQ} конвексна функција Δ^{mod}, што гарантује постојање јединственог минимума D^{MSPTQ}. Као и у случају SPTQ, да бисмо потврдили да итеративни поступак даје јединствену оптималну вредност за Δ^{mod}, у одељку са нумеричким резултатима (5.1.5), анализирамо резултате нумеричке оптимизације дисторзије по Δ^{mod}.

5.1.4 Примена два нова модела неуниформних квантизера у пост - тренинг квантизацији

Алгоритам дат у овом одељку илуструје процес наших експеримената: MNIST сетови података за обуку и тестирање се учитавају и затим преобликују у једнодимензионалне векторе од 784 (28x28) елемената. Свака компонента вектора је бинарна вредност која одређује интензитет пиксела. Наш NN модел (пре него што се примени квантизација) из разлога поређења је индентичан са онима наведеним у 4. поглављу [21] - [24]. Обука и процена тачности нашег NN модела је имплементирана у TensorFlow са Keras API, верзија 2.5.0 [76]. NN модел се састоји од 669 706 параметара тј. тежина, које се одређују током тренинга, а затим се квантују у овом одељку коришћењем нових двобитних неуниформних квантизера у пост-тренинг фази. Обука је спроведена у 10 епоха, док је број слика које се обрађују одједном (*batch size*) 128, што је резултирало са 469 итерација по епохи за завршетак обуке коришћењем 60 000 слика из сета за обуку. Тачност одређена на сету за валидацију након обуке износи 0.981, што значи да је NN модел направио тачна предвиђања за 98.1 % слика из сета за валидацију.

Као што је познато, различите дубине и архитектуре NN модела са потпуно повезаним слојевима би резултирали различитом тачношћу самог NN модела. Пре квантизације тежина, NN модел са много слојева и огромним бројем скривених неурона по слоју, може постићи бољу тачност. Међутим мањи модели NN се показују као бржи и при додавању нових слојева треба бити обазрив. За наведени NN модел, тзв. наш модел описан у радовима [21] и [24], обука NN модела, квантизација тежина и анализа тачности QNN су имплементирани у програмском језику Python [73]. У нашем QNN моделу, све тежине трениране NN су квантоване коришћењем једног од описаних нових неуниформних квантизера, да би се затим тачност нашег модела QNN и SQNR ових квантизера поредили за компримоване/квантоване тежине NN модела. Експерименталне перформансе SPTQ и MSPTQ се могу проценити одређивањем дисторзије или SQNR, који су дефинисани слично као у 4. поглављу и у раду [24]:

$$D_{\text{ex}}^{*}=\frac{1}{W}\left\|\hat{\boldsymbol{W}}-\hat{\boldsymbol{W}}^{*}\right\|_2^2=\frac{1}{W}\left(\hat{\boldsymbol{W}}-\hat{\boldsymbol{W}}^{*}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{W}}-\hat{\boldsymbol{W}}^{*}\right)=\frac{1}{W}\sum_{j=1}^{W}\left(w_j-w_j^{*}\right)^2, \tag{5.1.4.1}$$

$$\text{SQNR}_{\text{ex}}^{*}=10\log_{10}\left(\frac{\frac{1}{W}\sum_{j=1}^{W}w_j^2}{D_{\text{ex}}^{*}}\right), \tag{5.1.4.2}$$

где се * односи на примену SPTQ или MSPTQ. D_{ex}* и SQNR_{ex}* су експериментално одређена дисторзија и SQNR, $\hat{\boldsymbol{W}}=\{w_j\}_{j=1,2,\ldots,W}$, означава вектор тежина представљених у FP32 формату и $\hat{\boldsymbol{W}}^{*}=\{w_j^{*}\}_{j=1,2,\ldots,W}$, означава вектор квантованих тежина које треба учитати у QNN. Укратко, на самом почетку пост-тренинг квантизације, NN тежине се нормализују на нулту средњу вредност и јединичну варијансу чиме се формира вектор $\hat{\boldsymbol{W}}^{\mathbf{N}}=\{w_j^{\mathrm{N}}\}_{j=1,2,\ldots,W}$. Након што се све нормализоване тежине неуниформно квантују применом SPTQ или MSPTQ, и денормализују у оригинални опсег вредности тежина, тако формирани вектор квантованих тежина $\hat{W}^{*}=\{w_j^{*}\}_{j=1,2,\ldots,W}$ се учитава у QNN модел (видети алгоритам 5.1.4.1). Теоријски SQNR дефинишемо као:

$$\text{SQNR}_{\text{th}}^{\text{SPTQ}}=10\log_{10}\left(\frac{1}{D^{\text{SPTQ}}}\right), \tag{5.1.4.3}$$

$$\mathrm{SQNR}_{\mathrm{th}}^{\mathrm{MSPTQ}} = 10\log_{10}\left(\frac{1}{D^{\mathrm{MSPTQ}}}\right). \qquad (5.1.4.4)$$

који ће се бити упоређени са експериментално одређеним SQNR.

Алгоритам 5.1.4.1.- Компресија тежина у пост-тренинг квантизацији применом SPTQ / MSPTQ

Нотација: w_j –тежина трениране NN, w_j^{SPTQ} – квантована тежина применом SPTQ, w_j^{MSPTQ} – квантована тежина применом MSPTQ

Улаз: $\hat{W} = \{w_j\}_{j=1,\,2,\,\ldots,\,W,}$ тежине у FP32 формату, $\varepsilon_{\min}=10^{-4}$

Излаз: Квантоване тежине за SPTQ - $\hat{W}^{\mathrm{SPTQ}}=\{w_j^{\mathrm{SPTQ}}\}_{j\,=\,1,\,2,\,\ldots,\,W}$, квантоване тежине за MSPTQ - $\hat{W}^{\mathrm{MSPTQ}}=\{w_j^{\mathrm{MSPTQ}}\}_{j\,=\,1,\,2,\,\ldots,\,W}$, $\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{SPTQ}}$, $\mathrm{SQNR}_{\mathrm{th}}^{\mathrm{SPTQ}}$, $\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{MSPTQ}}$, $\mathrm{SQNR}_{\mathrm{th}}^{\mathrm{MSPTQ}}$, Тачност QNN

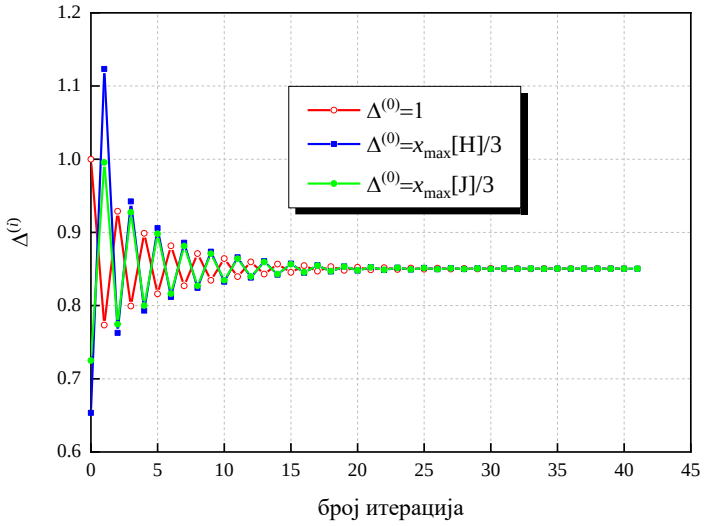
учитавање меморисаних тежина трениране NN: $\hat{W} = \{w_j\}_{j=1,\,2,\,\ldots,\,W}$
нормализација тежина и формирање вектора $\hat{W}^{\mathrm{N}} = \{w_j^{N}\}_{j=1,\,2,\,\ldots,\,W}$
$w_{\min}\leftarrow$ минимална вредност нормализованих тежина из $\hat{W}^{\mathrm{N}}$
$w_{\max} \leftarrow$ максимална вредност нормализованих тежина из $\hat{W}^{\mathrm{N}}$
<u>селекција SPTQ модела за квантизацију нормализованих тежина</u>
иницијализација $\varepsilon^{\mathrm{SPTQ}} \leftarrow 1$, $\Delta^{(0)} = \Delta^{\mathrm{SPTQ}} \leftarrow 1$ (или нека друга вредност), $i \leftarrow 1$
    while $\varepsilon^{\mathrm{SPTQ}} \geq \varepsilon_{\min}$ **do**
    срачунати $\Delta^{(i+1)}$ употребом (5.1.2.17)
    срачунати $\varepsilon^{\mathrm{SPTQ}} = \mathrm{abs}\,(\Delta^{(i+1)} - \Delta^{\mathrm{SPTQ}})$
    $\Delta^{\mathrm{SPTQ}} \leftarrow \Delta^{(i+1)}$
    $i \leftarrow i + 1$
    end while
$\Delta \leftarrow \Delta^{\mathrm{SPTQ}}$
$x_{\max}^{\mathrm{SPTQ}} \leftarrow 3\,\Delta$
срачунати $\{x_{-2}, x_{-1}, x_0, x_1, x_2\}$ употребом (5.1.2.7) за $x_{\max}^{\mathrm{SPTQ}}$
формирање кодне књиге $Y^{\mathrm{SPTQ}} = \{y_{-2}, y_{-1}, y_1, y_2\}$ употребом (5.1.2.8) или Табеле 5.1.3.1 а)
квантовање нормализованих тежина употребом кодне књиге Y^{SPTQ}
деномализација квантованих тежина и формирање вектора $\hat{W}^{\mathrm{SPTQ}}=\{w_j^{\mathrm{SPTQ}}\}_{j\,=\,1,\,2,\,\ldots,\,W}$
<u>селекција MSPTQ модела за квантизацију нормализованих тежина</u>
иницијализација $\varepsilon^{\mathrm{MSPTQ}} \leftarrow 1$, $\Delta^{\mathrm{mod}\,(0)} = \Delta^{\mathrm{MSPTQ}} \leftarrow \Delta^{\mathrm{SPTQ}}$, $i \leftarrow 1$
    while $\varepsilon^{\mathrm{MSPTQ}} \geq \varepsilon_{\min}$ **do**
    срачунати $\Delta^{\mathrm{mod}(i+1)}$ употребом (5.1.3.7)
    срачунати $\varepsilon^{\mathrm{MSPTQ}} = \mathrm{abs}\,(\Delta^{\mathrm{mod}(i+1)} - \Delta^{\mathrm{MSPTQ}})$
    $\Delta^{\mathrm{MSPTQ}} \leftarrow \Delta^{\mathrm{mod}(i+1)}$
    $i \leftarrow i + 1$
    end while
$\Delta^{\mathrm{mod}} \leftarrow \Delta^{\mathrm{MSPTQ}}$
$x_{\max}^{\mathrm{MSPTQ}} \leftarrow 3\,\Delta^{\mathrm{MSPTQ}}$
срачунати $\{x_{-2}^{\mathrm{mod}}, x_{-1}^{\mathrm{mod}}, x_0^{\mathrm{mod}}, x_1^{\mathrm{mod}}, x_2^{\mathrm{mod}}\}$ употребом Табеле 5.1.3.1 б) за $x_{\max}^{\mathrm{MSPTQ}}$
формирање кодне књиге $Y^{\mathrm{MSPTQ}} \equiv \{y_{-2}^{\mathrm{mod}}, y_{-1}^{\mathrm{mod}}, y_1^{\mathrm{mod}}, y_2^{\mathrm{mod}}\}$ употребом Табеле 5.1.3.1 б)
квантовање нормализованих тежина употребом кодне књиге Y^{MSPTQ}
деномализација квантованих тежина и формирање вектора $\hat{W}^{\mathrm{MSPTQ}}=\{w_j^{\mathrm{MSPTQ}}\}_{j\,=\,1,\,2,\,\ldots,\,W}$
срачунати $\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{SPTQ}}$, $\mathrm{SQNR}_{\mathrm{th}}^{\mathrm{SPTQ}}$, $\mathrm{SQNR}_{\mathrm{ex}}^{\mathrm{MSPTQ}}$, $\mathrm{SQNR}_{\mathrm{th}}^{\mathrm{MSPTQ}}$ употребом (5.1.2.14), (5.1.3.4), (5.1.4.1) - (5.1.4.4), процена тачности QNN.

5.1.5 Нумерички и експериментални резултати примене два нова модела неуниформних квантизера

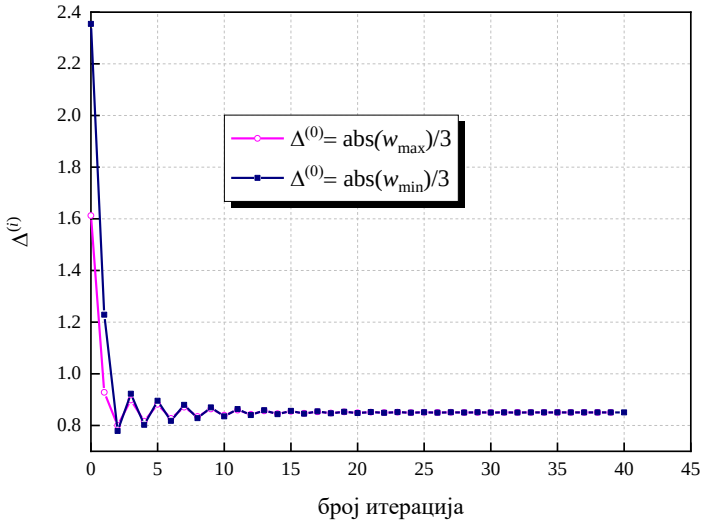
Позивајући се на алгоритам 5.1.4.1. за оба нова неуниформна квантизера описана у претходним одељцима, прво анализирамо број неопходних итерација за одређивање Δ и Δ^{mod}, или за одређивање x_{max}^{SPTQ} и x_{max}^{MSPTQ}. За иницијализацију алгоритма за одређивање Δ користимо различите вредности за $\Delta^{(0)}$ (које су наведене у табели 5.1.5.1.). За одређивање Δ^{mod} користимо резултат првог итеративног процеса, односно претпостављамо да је $\Delta^{mod(0)} = \Delta$. Као излазни критеријум алгоритма користимо исти услов, тј. да се две суседне итерације код одређивања Δ^{mod} разликују за мање од 10^{-4}. Посматрајући статистику нормализованих тежина нашег модела NN, утврдили смо да минимална и максимална нормализована тежина у FP32 формату пре квантизације износе $w_{min} = -7.063787$ и $w_{max} = 4.8371024$. Следећи унапред дефинисано правило за одређивање гранулаарног региона као $[-3\Delta, 3\Delta]$ за SPTQ, за иницијализацију алгоритма за SPTQ користимо $\Delta^{(0)} = |w_{max}|/3 = 1.61237$ и $\Delta^{(0)} = |w_{min}|/3 = 2.3546$. Такође разматрамо и $\Delta^{(0)} = x_{max}[H] /3 = 0.6536$ као и $\Delta^{(0)} = x_{max}[J] /3 = 0.7249$, где су $x_{max}[H]$ и $x_{max}[J]$ оптималне и асимптотски оптималне x_{max} вредности за двобитни униформни квантизер које су одредили *Hui* [62] и *Jayant* [2] (видети табелу 5.1.5.1.). Различите иницијализације захтевају око 40 итерација (видети табелу 5.1.5.1 и слику 5.1.5.1). Штавише, треба приметити да све уочене иницијализације доводе до јединствене коначне вредности Δ ($\Delta = 0.8504$) и $x_{max}^{SPTQ} = 3\Delta = 2.5512$. Ако даље користимо $\Delta^{mod(0)} = \Delta = 0.8504$ за итеративно одређивање Δ^{mod}, с обзиром на исти излазни критеријум алгоритма, потребно нам је само 7 итерација. Као резултат другог итеративног процеса за MSPTQ, одређујемо $\Delta^{mod} = 0.9021$ као и $x_{max}^{MSPTQ} = 3\Delta^{mod} = 2.7063$.

Табела 5.1.5.1. Број итерација за одређивање x_{max}^{SPTQ} за различите иницијализације.

SPTQ	$\Delta^{(0)} = 1$	$\Delta^{(0)} = \lvert w_{max}\rvert/3$	$\Delta^{(0)} = \lvert w_{min}\rvert/3$	$\Delta^{(0)} = x_{max}[H] /3$	$\Delta^{(0)} = x_{max}[J] /3$
број итерација	40	39	40	41	41



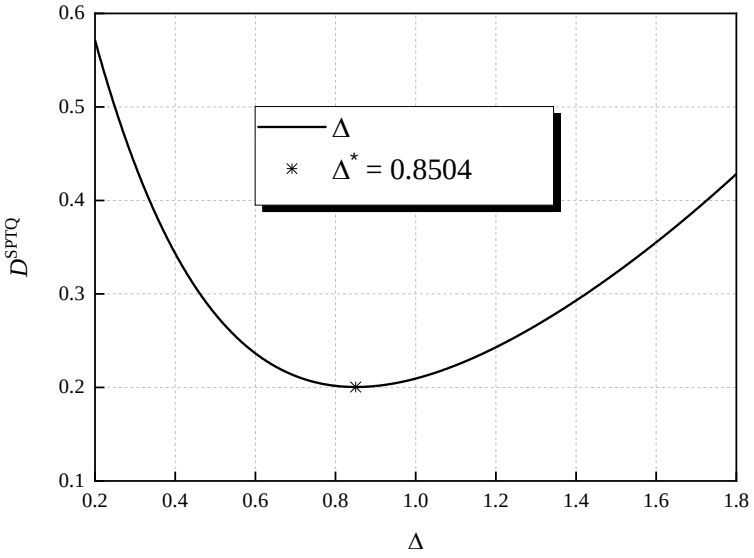
а)



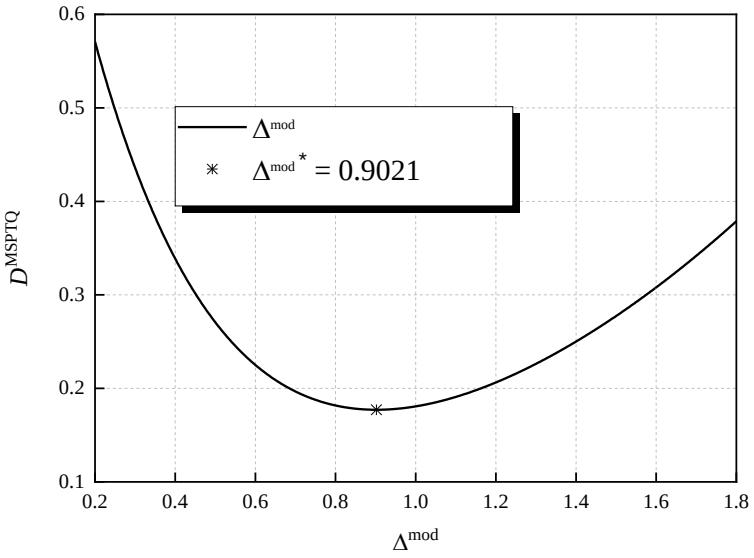
б)

Слика 5.1.5.1. Илустрација конвергенције алгоритма 5.1.4.1. за

а) $\Delta^{(0)}=1$, $\Delta^{(0)}=x_{max}[H]/3$, $\Delta^{(0)}=x_{max}[J]/3$; **б**) $\Delta^{(0)}=|w_{max}|/3$, $\Delta^{(0)}=|w_{min}|/3$.



а)



б)

Слика 5.1.5.2. Зависност дисторзије од иницијалног квантизационог корака за а) SPTQ, б) MSPTQ.

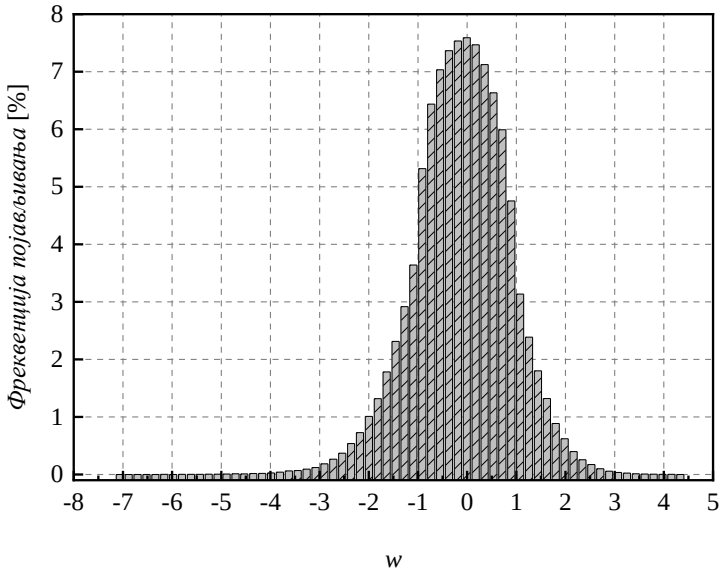
Да бисмо додатно потврдили да смо излазним критеријумом, специфицираним у алгоритму 5.1.4.1, добили оптималне вредности за Δ и Δ^{mod}, на слици 5.1.5.2. приказујемо зависности дисторзије примењених квантизера од одговарајућих основних величина корака Δ и Δ^{mod}. Итеративно добијене крајње вредности за Δ and Δ^{mod} су означене звездицама на слици 5.1.5.2. и заиста су оптималне, јер дају минимум D^{SPTQ} и D^{MSPTQ}. Даље у овом одељку представљамо експериментално добијене резултате примене SPTQ и MSPTQ у пост-тренинг квантизацији тежина NN модела. Да бисмо анализирали перформансе описаних квантизера у квантизацији тежина NN, спроводимо експерименте за различите специфичне изборе грануларног региона за неуниформне квантизере, који имају за циљ да пруже увид у утицај различитих ширина грануларног региона на перформансе неуниформних квантизера и тачност QNN.

Грануларни регион у нашем случају 1 је дефинисано као $[-\min(|w_{min}|,|w_{max}|), \min(|w_{min}|,|w_{max}|)]$, што је у нашем експерименту једноставно $[-w_{max}, w_{max}]$. Према томе, у случају 1, ширина грануларног региона зависи од максималне вредности нормализованих тежина обученог NN модела, што за посматране нормализоване тежине износи $w_{max} = 4.8371024$. Одређивањем грануларног региона за SPTQ и MSPTQ као што је наведено у случају 1, обухвата се 99.988% свих нормализованих тежина.

Из табеле 5.1.5.2. се може приметити да овако дефинисан случај 1 пружа највећу тачност QNN од свих посматраних случајева, која износи 97.61%.

Табела 5.1.5.2. SQNR и тачност QNN при примени различитих двобитних SPTQ.

$w_{min} = -7.063787$, $w_{max} = 4.8371024$, $3\Delta = 2.5512$ $x_{max}[H] = 1.9605$, $x_{max}[J] = 2.1748$	Случај 1 $\Re_g$ $[-w_{max}, w_{max}]$	Случај 2 $\Re_g$ $[w_{min}, -w_{min}]$	Случај 3 $\Re_g$ $[-3\Delta, 3\Delta]$	Случај 4 $\Re_g$ $[-x_{max}[H], x_{max}[H]]$	Случај 5 $\Re_g$ $[-x_{max}[J], x_{max}[J]]$
$SQNR_{ex}^{SPTQ}$(dB)	4.6899	2.5745	7.8099	7.9051	**8.0068**
$SQNR_{th}^{SPTQ}$(dB)	4.4438	1.6044	**6.9790**	6.5437	6.8086
Тачност (%)	**97.61**	97.42	95.75	93.77	94.52
Припадност $\Re_g$ (%)	99.988	100	98.567	94.787	96.691



Слика 5.1.5.3. Нормализовани хистограм нормализованих тежина (FP32 формат) трениранoг модела NN за MNIST сет података.

Компарацијом са случајем у коме је примењен једноставни униформни квантизер, можемо закључити да применом SPTQ за квантовање нормализованих тежина обезбеђујемо повећање тачности од 0.64% (видети табелу 5.1.5.3.). Ово повећање тачности је значајно, посебно узимајући у обзир да је једина разлика у примењеним квантизерима док се претпоставља иста битска брзина. Слично, експериментални и теоријски добијени SQNR за SPTQ је већи у поређењу са униформним квантизером, уз пораст SQNR вредности од 1.8078 dB и 2.5078 dB, респективно. Можемо приметити да теоријски утврђени SQNR има ниже вредности од експериментално остваренog SQNR. Као што је објашњено у 4. поглављу, као и у радовима [21], [24], разлог је исти и базира се на чињеници да се у експерименталној анализи квантују тежине који потичу из расподеле сличне Лапласовој које узимају вредности из ограниченог скупа могућих вредности [−7.063787, 4.8371024] (видети слику 5.1.5.3.), док се у теоријској анализи претпоставља квантизација неограничене Лапласове функције густине вероватноће, што условљава повећање дисторзије, односно смањење теоријске вредности SQNR. Укратко, случај 1 већ показује предности имплементације SPTQ у

односу на униформни квантизер обезбеђујући повећање свих значајних перформансних индикатора.

У случају 2, грануларни регион је $[-\max(|w_{\min}|, |w_{\max}|), \max(|w_{\min}|, |w_{\max}|)]$ и у нашем експерименту се може изразити као $[-|w_{\min}|, |w_{\min}|]$, што у пракси постаје $[w_{\min}, -w_{\min}]$, те је грануларни регион [−7.063787, 7.063787]. Може се приметити да грануларни регион у овом случају обухвата 100 % тежина и чак превазилази максималну вредност нормализованих тежина, што га чини непотребно широким и репрезентативним за неповољан избор $\Re_g$. Дакле, у овом случају добијамо најнижу уочену вредност SQNR међу свим разматраним случајевима, али је она значајно већа од вредност SQNR добијене применом униформног квантизера, која постиже чак и негативне вредности. За разлику од ниске вредности SQNR, у случају 2, тачност QNN је веома висока са постигнутих 97.42 %, а посебно ако се има у виду да за униформни квантизер и исти грануларни регион, тачност износи само 94.58 %. Ово запажање наглашава предности коришћења неуниформне квантизације, која обезбеђује боље перформансе QNN, чак и у случају избора прешироког грануларног региона.

У случају 3, грануларни регион, $\Re_g$, одређен је за итеративно израчунати оптимални корак квантизације $\Delta = 0.8504$. Како овде прво разматрамо SPTQ, ширина грануларног региона је дефинисана као $x_{\max}^{SPTQ} = 3\Delta = 2.5512$. Ова вредност представља теоријски оптималну вредност ширине грануларног региона, чиме се обезбеђује максимална теоријска вредност SQNR за разматрани случај.

Табела 5.1.5.3. SQNR и тачност QNN при примени различитих двобитних униформних квантизера (део резултата из [21]).

$w_{\min} = -7.063787$, $w_{\max} = 4.8371024$, $3\Delta = 2.5512$ $x_{\max}[H] = 1.9605$, $x_{\max}[J] = 2.1748$	Случај 1 $\Re_g$ $[-w_{\max}, w_{\max}]$	Случај 2 $\Re_g$ $[w_{\min}, -w_{\min}]$	Случај 3 $\Re_g$ $[-3\Delta, 3\Delta]$	Случај 4 $\Re_g$ $[-x_{\max}[H], x_{\max}[H]]$	Случај 5 $\Re_g$ $[-x_{\max}[J], x_{\max}[J]]$
$SQNR_{ex}^{UQ}$(dB)	2.8821	-1.2402	8.2325	**8.7676**	8.7639
$SQNR_{th}^{UQ}$(dB)	1.9360	-2.0066	6.8237	6.9787	**7.0707**
Тачност (%)	96.97	94.58	**97.12**	96.34	96.74
Припадност $\Re_g$ (%)	99.988	**100**	98.567	94.787	96.691

Иако случај 3 даје највиши теоријски и виши експериментални SQNR у поређењу са претходним случајевима 1 и 2, тачност QNN је знатно нижа. Ово даље потврђује премису да квантизер који даје највећи SQNR не обезбеђује нужно највећу тачност QNN за дати избор $\Re_g$.

Случајеви 4 и 5 користе добро познате, оптималне и асимптотски оптималне вредности ширине грануларног региона за двобитни униформни квантизер из литературе, које су одредили *Hui* [62] и *Jayant* [2]. Иако ове вредности нису оптималне за наш SPTQ, ми их укључујемо у анализу да бисмо стекли бољи увид у утицај избора $\Re_g$ на оба перформансна индикатора - SQNR и тачност QNN. Може се приметити да иако нису оптимални за SPTQ, величине грануларног региона наведене у случајевима 4 и 5 дају највише експерименталне вредности SQNR, при чему је теоријски SQNR релативно близу оптималној вредности са максималном разликом од око 0.4 dB. За разлику од вредности SQNR, у ова два случаја добијамо најнижу тачност QNN. Нижа тачност је директан резултат преуског грануларног региона, $\Re_g$, са 94.787 % и 96.691 % тежина унутар $\Re_g$ за случајеве 4 и 5, респективно.

Анализом свих уочених случајева може се закључити да $\Re_g$ има релативно велики утицај на тачност QNN и остварени SQNR. Такође, показало се да је SPTQ много бољи од униформног квантизера за случај ширег $\Re_g$, што смо интуитивно и очекивали, а такође је и опште је познато у неуниформној квантизацији [2]. Насупрот овоме, може се приметити да тачност QNN са применом униформног квантизера надмашује тачност QNN за SPTQ у случајевима 3, 4 и 5, при чему је случај 3 најзначајнији, јер представља оптимално решење за теоријски $\mathrm{SQNR}^{\mathrm{SPTQ}}$. На основу SQNR анализе за SPTQ, можемо закључити да би SPTQ дизајниран као у случају 3 могао бити у великој мери применљив у традиционалним задацима квантизације код обраде и преноса сигнала. Да бисмо превазишли поменуту несавршеност SPTQ у квантизацији неуронских мрежа, као што је већ описано, разматрамо једноставан и ефикасан модел MSPTQ, чије перформансе у наставку детаљно анализирамо.

Табела 5.1.5.4. SQNR и тачност QNN при примени различитих двобитних MSPTQ за квантовање нормализованих тежина.

$w_{min} = -7.063787$, $w_{max} = 4.8371024$, $x_{max}^{SPTQ} = 3\Delta = 2.5512$ $x_{max}^{MSPTQ} = 3\Delta^{mod} = 2.7063$	Случај 1 $\Re_g$ $[-w_{max}, w_{max}]$	Случај 2 $\Re_g$ $[w_{min}, -w_{min}]$	Случај 3 $\Re_g$ $[-3\Delta, 3\Delta]$	Случај 4 $\Re_g$ $[-3\Delta^{mod}, 3\Delta^{mod}]$
$SQNR_{ex}^{MSPTQ}$(dB)	5.5741	2.9114	**8.6839**	8.5608
$SQNR_{th}^{MSPTQ}$(dB)	5.0581	1.9158	7.4890	**7.5165**
Тачност (%)	**97.91**	96.98	97.17	97.23
Припадност $\Re_g$ (%)	99.988	100	98.567	99.001

Дефинисали смо да MSPTQ уводи једноставну модификацију у специфицирању границе одлучивања како би се побољшале перформансе SPTQ за случајеве ужег $\Re_g$, укључујући онај који је конструисан са оптималном вредношћу грануларног региона. Перформансе MSPTQ, представљене су у табели 5.1.5.4. за мало другачије случајеве грануларног региона. Као што је претходно поменуто, случајеви 4 и 5 нису релевантни за неуниформно квантовање. Случајеви 1, 2 и 3 су исти као у претходној анализи, тако да можемо спровести директно поређење перформанси. Приметно је да MSPTQ надмашује и SPTQ и униформни квантизер у свим посматраним перформансним индикаторима за случајеве 1, 2 и 3, док је тачност QNN са применом MSPTQ већа за 0.44 % у случају 2. Штавише, MSPTQ остварује добитак и теоријске и експерименталне вредности SQNR, за све компаративне случајеве модела QNN са имплементираним униформним квантизером за квантизаацију нормализованих тежина.

Случај 2 је специфичан, јер је пример неповољног избора $\Re_g$, а QNN са применом SPTQ остварује већу тачност у односу на MSPTQ. С друге стране, за пажљиво одабрану ширину $\Re_g$, применом MSPTQ остварујемо значајно повећање и SQNR и тачности QNN у поређењу са SPTQ. Добитак у тачности је посебно значајан у случају 3, који користи грануларни регион оптимално одређен за SPTQ, износи 1.42 % (видети табелу 5.1.5.2.). За $\Re_g$ специфициран у случају 3, MSPTQ постиже највећу

експерименталну вредност SQNR, са разликом од 0.874 dB у поређењу са SPTQ и 0.4514 dB у поређењу са униформним квантизером.

Случај 4 користи нумерички оптимизовану ширину грануларног региона, $\Re_g$ за MSPTQ, при чему је вредност грануларног региона блиска оној одређеној за SPTQ. Очекивано, теоријски одређени SQNR је највећи за овај случај, док експериментално остварени SQNR, као и тачност QNN, имају другу највећу вредност међу свим случајевима, са 99.001 % нормализованих тежина унутар $\Re_g$. Како деградација тачности у случају 4 износи око 0.9 % у поређењу са референтном тачношћу нашег модела пре квантизације (98.1 % − 97.23 % = 0.87 %), можемо закључити да смо једноставном модификацијом успели не само да побољшамо SQNR, већ и да повећамо тачност QNN у поређењу са случајем где су имплементирани униформни квантизер и SPTQ.

6. Закључак

У овој дисертацији су описани нови модели скаларних квантизера, чије су перформансе посебно анализиране и упоређене са квантизерима сличне сложености, али не и дизајна. Такође је спровођена и оптимизација описаних модела, како би се њихове перформансе додатно побољшале. У области обраде сигнала смо се ослањали на математичке моделе и усвојену Лапласову функцију густине вероватноће, што је дало једноставнији увид у процену перформанси, која се базира на одређивању SQNR. У дисертацији су дати нови модели квантизера, као што је PWUQ за Лапласову функцију густине расподеле амплитуда одмерака сигнала на улазу квантизера, за који је предложена нова идеја за поделу амплитудског опсега квантизера на два региона, CGR и PGR. Ширине ова два региона су оптимизоване коришћењем итеративног алгоритма, који задовољава услов минималне дисторзије PWUQ модела. У самом PWUQ моделу су примењени најједноставнији униформни квантизери са једнаким битским брзинама у оба региона, што дизајн овог модела чини много једноставнијим у поређењу са многим неуниформним моделима квантизера доступним у литератури. Најзначајнији допринос се огледа у SQNR добитку оствареном у односу на униформни квантизер, што даље оправдава смисленост описаног модела.

Такође смо предложили и описали модел симетричног квантиле квантизера и дефинисали скуп параметара који описују SQQ. Описани SQQ модел користи погодности компандора за директно рачунање нивоа одлуке, док су репрезентациони нивои оптимално дефинисани, чиме се прави основна разлика у односу на модел компандора где су и нивои одлучивања и репрезентациони нивои одређени у складу са специфичном функцијом компресора.

Пратећи главни циљ кодовања и компресије сигнала, а то је смањење брзине преноса уз њено усклађивање са апликативним захтевима, за неуниформне изворе, као што је Лапласов, који је претпостављен у анализама у овој дисертацији, неуниформна квантизација омогућава боље коришћење доступне битске брзине, те је предложен и

описан начин пројектовања квазилогаритамског квантизера. Како је сам квазилогаритамски квантизер робустан у широком динамичком опсегу варијансе сигнала, препознали смо посебан интерес да у зависности од коефицијента скалирања, који даје однос варијансе на улазу у квантизер и варијансе за коју је квантизер пројектован, одредимо максимални SQNR за различите битске брзине, као и за различите вредности фактора компресије μ. Од посебног значаја је итеративно одређивање амплитуде максималног оптерећења квазилогаритамског квантизера, те су вршене различите иницијализације итеративног метода, где смо указали на брзину конвергенције итеративног процеса.

Како су квантизери нашли примену у атрактивној области NN, која је мање осетљива на фину параметризацију квантизера у односу на обраду сигнала, проверили смо најпре модалитете примене нискобитних униформних квантизера на усвојеном моделу NN, где је квантизација примењивана у пост-тренинг фази. Сам трослојни FC модел NN је коришћен у овој дисертацији као показни модел NN, обзиром да смо хтели да изолујемо утицај примене различитих модела нискобитних скаларних квантизера на перформансе QNN, примарно тачност QNN и експериментални SQNR квантизера.

Главни закључци до којих смо дошли су даље резимирани:

- Нисмо уочили једноставну и једнозначну релацију између квалитативних перформансних индикатора, SQNR (теоријски и експериментални) и тачности QNN, при примени различитих модела скаларних квантизера. Потврдили смо премису да квантизер који обезбеђује највећи SQNR (теоријски или експериментални) не мора нужно да обезбеди највећу тачност QNN.
- Показали смо да смо коришћењем двобитног униформног квантизера за компресију тежина модела NN успели значајно да сачувамо тачност модела QNN, која је деградирана за 1.13 % у односу на референтну тачност модела NN (пре примене квантизације).
- Утврдили смо да двобитна униформна квантизација по слојевима може да обезбеди додатно побољшање тачности описаног модела QNN за MNIST скуп

података. Приликом прилагођавања по слојевима, приметили смо да расподеле тежина варирају у различитим слојевима модела NN, при чему у случају описаног модела NN, трећи слој показује највеће одступање од Лапласове функције густине вероватноће. То запажање отвара могућност адаптације по слојевима, не само за $\Re_g$, већ и за функцију густине вероватноће која се користи за моделовање тежина NN модела у различитим слојевима. Док је величина корака униформног квантизера искључиво дефинисана помоћу ширине $\Re_g$, таква адаптација би могла бити посебно корисна када се користи неуниформна квантизација, где би се величина корака прилагођавала расподели улазних података.

- Како смо код двобитног униформног квантизера показали да је одређивање ширине $\Re_g$ јако битно, сличну анализу утицаја избора ширине $\Re_g$ на перформансне индикаторе смо поновили и за тробитни униформни квантизер. Анализа је спроведена за MNIST и FASHION-MNIST скуп података, као и за MLP и CNN. Утврдили смо да се за брзину од R = 3 бит/одмерак утицај избора $\Re_g$ на тачност QNN значајно смањује у поређењу са случајем где су идентичне тежине сачуване у FP32 формату и униформно квантоване брзином од 2 бит/одмерак.
- Испитивали смо приоритет примене квантизације тежина слојева нашег модела NN, што нас је довело до закључка, да је најважнији слој за квантизацију тежина управо први слој NN, са највећим бројем тежина, док је последњи слој, иако садржи најмањи број тежина, други по важности слој за примену квантизације тежина. Такође смо илустровали и расподелу нормализованих тежина последњег слоја, која највише одступа од Лапласове расподеле и најприближнија је униформној расподели. Закључак изведен за приоритет слојева модела NN за квантизацију би могао бити посебно користан за прелиминарну примену квантизације тежина у већим NN архитектурама, уз примену квантизације мешовите прецизности, тј. битске брзине.
- У поглављу посвећеном неуниформним квантизерима предложили смо два нова модела двобитог неуниформног квантизера, SPTQ и MSPTQ, који користе

једно од две особине најједноставнијег униформног квантизера, и испитали смо да ли су погоднији за примену у пост- тренинг квантизацији од униформног квантизера. Оптимизацијом дисторзије описаних неуниформнин квантизера, а да бисмо постигли највиши теоријски SQNR, извели смо формуле за итеративно одређивање основних величина корака оба неуниформна квантизера, Δ и Δ^{mod}, што је од највеће важности у традиционалној квантизацији. Доказали смо да су коришћени итеративни алгоритми дали вредности Δ и Δ^{mod} које су заиста оптималне, јер се поклапају са одговарајућим резултатима нумеричке оптимизације дисторзије по основним величинама корака и дају минимум дисторзије за SPTQ и MSPTQ. Увођењем MSPTQ модела успели смо не само да побољшамо SQNR, већ и да повећамо тачност QNN, у поређењу са случајем где су имплементирани двобитни униформни квантизер и SPTQ.

Иако је униформни квантизер веома искоришћен модел квантизера, због своје интригантне природе и могућности да се донекле модификује, природно је очекивати да ће истраживање и развој неких модификованих модела квантизера наставити да привлачи пажњу научне заједнице.

Предложени и описани модели квантизера дати у овој дисертацији су значајни за области примене: обраду сигнала и квантоване неуронске мреже. Детаљни описи модела квантизера, као и њихова оптимизација су публиковани у часописима са импакт фактором и наведени су у референцама.

7. Литература

[1] A. Kondoz, Digital Speech: Coding for Low Bit Rate Communication Systems, John Wiley & Sons, 2nd edition, October 20, 2004.

[2] N. S. Jayant, P. Noll, Digital Coding of Waveforms, Prentice-Hall, New Jersey, 1984.

[3] D. Salomon, Data Compression – The Complete Reference, Springer, 4th edition, 2007.

[4] K. Sayood, Introduction to Data Compression, Morgan Kaufmann, 5th edition, 2018.

[5] L. Hanzo, C. Somerville, J. Woodard, Voice and Audio Compression for Wireless Communications, John Wiley & Sons - IEEE Press, 2nd edition, October 8, 2007.

[6] W. A. Pearlman, A. Said, Digital Signal Compression: Principles and Practice, Cambridge University Press, 2011.

[7] G. K. Wallace, "The JPEG Still Picture Compression Standard," *IEEE Transactions on Consumer Electronics*, vol. 38 (1), pp. 18-34, 1992.

[8] A. Gersho, R. M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Masachusetts, 1992.

[9] A. Gersho, "Quantization," *IEEE Communications Society Magazine*, pp. 20-29, September 1977.

[10] W.R. Bennett, "Spectra of Quantized Signals," *The Bell Systems Technical Journal*, vol. 27, pp. 446-472, July 1948.

[11] Widrow, I. Kollar, Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications, Cambridge University Press, 1st edition, July 21, 2008.

[12] S. Na, D. Neuhoff, "Monotonicity of Step Sizes of MSE-Optimal Symmetric Uniform Scalar Quantizers," *IEEE Transactions on Information Theory 2019*, vol. 65, pp. 1782-1792, 2019.

[13] S. Na, "On the Support of Fixed-Rate Minimum Mean-Squared Error Scalar Quantizers for a Laplacian Source," *IEEE Transactions on Information Theory*, vol. 50 (5), pp. 937-944, 2004.

[14] S. Na, D. Neuhoff, "On the Convexity of the MSE Distortion of Symmetric Uniform Scalar Quantization," *IEEE Transactions on Information Theory*, vol. 64, pp. 2626-2638, 2018.

[15] Guo, Y. "A Survey on Methods and Theories of Quantized Neural Networks", arXiv 2018.

[16] A. Gholami, S. Kim, Z. Dong, Z. Yao, M.W. Mahoney, K. Keutzer, "A Survey of Quantization Methods for Efficient Neural Network Inference", arXiv 2021.

[17] J. Nikolić, D. Aleksić, Z. Perić, M. Dinčić, "Iterative Algorithm for Parameterization of Two-Region Piecewise Uniform Quantizer for the Laplacian Source", *Mathematics*, vol. 9(23), 3091, 2021.

[18] Z. Perić, J. Nikolić, D. Aleksić, A. Perić, "Symmetric Quantile Quantizer Parameterization for the Laplacian Source: Qualification for Contemporary Quantization Solutions", *Mathematical Problems in Engineering*, vol. 2021, Article ID 6647135, 12 pages, 2021.

[19] D. Aleksic, Z. Perić, "Analysis and design of Robust Quasilogarithmic Quantizer for the Purpose of Traffic Optimization", *Information Technology and Control*, vol. 47(4), pp. 615-622, 2018.

[20] Z. Perić, D. Aleksić, "Quasilogarithmic Quantizer for Laplacian Source: Support Region Ubiquitous Optimization Task", *Revue Roumaine des Sciences Techniques - Électrotechn. et Énergetique,* vol. 64 (4), pp. 403-408, 2019.

[21] S. Tomić, J. Nikolić, Z. Perić, Danijela Aleksić, "Performance of Post-Training Two-Bits Uniform and Layer-Wise Uniform Quantization for MNIST Dataset from the Perspective of Support Region Choice", *Mathematical Problems in Engineering*, vol. 2022, Article ID 1463094, 2022.

[22] J. Nikolić, Z. Perić, S. Tomić, D. Aleksić, "On Different Criteria for Optimizing the Two-bit Uniform Quantizer", *INFOTEH-JAHORINA* 2022, Conference Proceedings, Jahorina, RS, BiH, March 16-18, 2022.

[23] J. Nikolić, S. Tomić, Z. Perić, D. Aleksić "Analysis of Neural Network Accuracy Degradation due to Uniform Weight Quantization of One or More Layers", *57th*

International Conference on Information, Communication and Energy Systems and Technologies, accepted for ICEST 2022, June 16-18, 2022.

[24] J. Nikolić, Z. Perić, D. Aleksić, S. Tomić, A. Jovanović, "Whether the Support Region of Three-bit Uniform Quantizer has a Strong Impact on Post-training Quantization for MNIST Dataset? ", *Entropy*, vol. 23 (12), 1699, 2021.

[25] R. M. Gray, D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44 (6), pp. 2325-2383, 1998.

[26] C. Shannon, "A Mathematical Theory of Communication", *The Bell System Technical Journal*, vol. 27 (3), pp. 379-423, 1948.

[27] B. Oliver, J. Pierce, C. Shannon. "The Philosophy of PCM", *Proceedings of the IRE*, vol. 36 (11), pp. 1324-1331, 1948.

[28] S. Gazor, W. Zhang, "Speech Probability Distribution," *IEEE Signal Processing Letters*, vol. 10 (7), pp. 204-207, 2003.

[29] P. Kabal, "Quantizers for the Gamma Distribution and Other Symmetrical Distributions", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32(4), pp. 836-841, 1984.

[30] J. Lee, S. Na, "A Rigorous Revisit to the Partial Distortion Theorem in the Case of a Laplacian Source," *IEEE Communications Letters*, vol. 21 (12), pp. 2554-2557, 2017.

[31] S. Kotz, T. Kozubowski, K. Podgórski, "The Laplace Distribution and Generalization: A Revisit with Applications to Communications, Economics, Engineering, and Finance, Springer Science & Business Media", 2001.

[32] S. Naik, R. Jagannath, V. Kuppili, "Bat Algorithm-based Weighted Laplacian Probabilistic Neural Network", *Neural Computing and Applications*, vol. 32 (4), pp. 1157-1171, 2020.

[33] V. Delić, Z. Perić, M. Sečujski, N. Jakovljević, J. Nikolić, D. Mišković, N. Simić, S. Suzić, T. Delić, "Speech Technology Progress based on new Machine Learning Paradigm", *Compuational Intelligence and Neuroscience*, Article ID 4368036, 19 pages, 2019.

[34] Z. Perić, M. Petković, J. Nikolić, "Optimization of Multiple Region Quantizer for Laplacian Source", *Digital Signal Processing*, vol. 27, pp. 150-158, 2014.

[35] X. Cui, X. Li and B. Wang, “Communication Optimization Technology based on Network Dynamic Performance Model”, *Mathematical Problems in Engineering*, vol. 2020, Article ID 8890721, 13 pages, 2020.

[36] N. Strivastava, G. Hinton, A. Krizhevsky et al., “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, *Journal of Machine Learning Research*, vol. 15 (1), pp. 1929-1958, 2014.

[37] X. Long, X. Zeng, Z. Ben, D. Zhou, M. Zhang, “A Novel Low-Bit Quantization Strategy for Compressing Deep Neural Networks”, *Compuational Intelligence and Neuroscience*, 2020.

[38] I. Hubara, M. Courbariaux, D. Soudry, E. Y. Ran, Y. Bengio, “Binarized Neural Networks”, *Proceedings of the 30th Conference on Neural Information Processing Systems* (*NeurIPS 2016*), Barcelona, Spain, 1-9. December 2016.

[39] D. Lin, S. Talathi, D. Soudry, S. Annapureddy, “Fixed Point Quantization of Deep Convolutional Networks”, *Proceedings of the 33rd International Conference on Machine Learning Conference on Neural Information Processing Systems*, New York, NY, USA, 2849-2858, June 2016.

[40] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, Y. Bengio, “Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations”, *Journal of Machine Learning Research*, vol. 18, pp. 6869-6898, 2017.

[41] K. Huang, B. Ni, D. Yang, “Efficient Quantization for Neural Networks with Binary Weights and Low Bit Width Activations”, *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 3854-3861, 2019.

[42] Z. Yang, Y. Wang, K. Han, C. Xu, C. Xu, D. Tao, C. Xu, “Searching for Low-Bit Weights in Quantized Neural Networks”, *Proceedings of the 34th Conference on Neural Information Processing Systems* (*NeurIPS 2020*), Vancouver, Canada, 6-12. December 2020.

[43] M. Vestias, R. Duarte, J. Sousa, H. Neto, “Moving Deep Learning to the Edge”, *Algorithms*, vol. 13 (5), 125, 2020.

[44] S. Uhlich, L. Mauch, F. Cardinaux, K. Yoshiyama, “Mixed Precision DNNs: All You Need is a Good Parametrization”, *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, April 2020.

[45] Z. Perić, B Denić, M. Savić, N. Vučić, N. Simić, “Binary Quantization Analysis of Neural Networks Weights on MNIST Dataset”, *Elektronika Ir Elektrotechnika*, vol. 27 (4), pp. 55-61, 2021.

[46] D. Liu, H. Kong, X. Luo, W. Liu, R. Subramaniam, “Bringing AI to edge: From Deep Learning's Perspective”, arXiv 2020.

[47] T. Salimans, D. Kingma, “Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks”, arXiv 2018, arXiv:1602.07868.

[48] S. Sanghyun, K. Juntae, “Efficient Weights Quantization of Convolutional Neural Networks Using Kernel Density Estimation Based Non-Uniform Quantizer”, *Applied Sciences*, vol. 9 (12), 2559, 2019.

[49] Z. Perić, B. Denić, M. Dinčić, J. Nikolić, “Robust 2-bit Quantization of Weights in Neural Network Modeled by Laplacian Distribution”, *Advances in Electrical and Computer Engineering*, vol. 21, pp. 3-10, 2021.

[50] R. Banner, Y. Nahshan and D. Soudry, “Postraining 4-bit quantization of convolutional networks for rapid-deployment”, *Proceedings of the 33rd Conference on Neural Information Processing Systems* (*NeurIPS 2019*), Vancouver, Canada, pp. 7948-7956, 2019.

[51] H. Pham, M. Guan, B. Zoph, Q. Le, J. Dean, “Efficient Neural Architecture Search via Parameters Sharing”, *Proceedings of the International Conference on Machine Learning*, pp. 4095-4104, 2018.

[52] H. Cai, C. Gan, T. Wang, Z. Zhang, S. Han, “Once-for-all: Train One Network and Specialize it for Efficient Deployment”, arXiv 2019.

[53] X. Xiao, Z. Wang, S. Rajasekaran, “Autoprune: Automatic Network Pruning by Regularizing Auxiliary Parameters”, *Advances in Neural Information Processing Systems*, pages 13681-13691, 2019.

[54] R. Yu, A. Li, C. F. Chen, J. H. Lai, V. Morariu, X. Han, M. Gao, C. Y. Lin, L. S. Davis, “Nisp: Pruning Networks Using Neuron Importance Score Propagation,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9194-9203, 2018.

[55] G. Hinton, O. Vinyals, J. Dean, “Distilling the Knowledge in a Neural Network”, arXiv preprint arXiv:1503.02531, 2015.

[56] J. Tee, D. P. Taylor, “Is Information in the Brain Represented in Continuous or Discrete Form?,” *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 6 (3), pp. 199–209, 2020.

[57] S. Na, D. Neuhoff, “On the Support of MSE-Optimal, Fixed-Rate, Scalar Quantizers,” *IEEE Transactions on Information Theory*, vol. 47 (7), pp. 2972-2982, 2001.

[58] S. Na, “Asymptotic Formulas for Mismatched Fixed-Rate Minimum MSE Laplacian Quantizers,” *IEEE Signal Processing Letters*, vol. 15, pp. 13-16, 2008.

[59] Z. Perić, J. Nikolić, “An Effective Method for Initialization of Lloyd-max’s Algorithm of Optimal Scalar Quantization for Laplacian Source”, *Informatica*, vol. 18 (2), pp. 279-288, 2007.

[60] J. Nikolić, Z. Perić, “Lloyd-Max’s Algorithm Implementation in Speech Coding Algorithm Based on Forward Adaptive Technique”, *Informatica*, vol. 19 (2), pp. 255-270, 2008.

[61] P. Panter, W. Dite, “Quantization Distortion in Pulse Count Modulation with Nonuniform Spacing of Levels”, *Proceedings* IRE, vol. 39, pp. 44-48, 1951.

[62] D. Hui, D. Neuhoff, “Asymptotic Analysis of Optimal Fixed-Rate Uniform Scalar Quantization,” *IEEE Transactions on Information Theory*, vol. 47 (3), pp. 957-977, 2001.

[63] M. Tančić, Z. Perić, N. Simić, S. Tomić, “Performance of the Quasi-Logarithmic Quantizer for Discrete Input Signals”, *Information Technology and Control*, vol. 46 (3), pp. 395-402, 2017.

[64] S. Tomić, Z. Perić, M. Tančić, J. Nikolić, “Backward Adaptive and Quasi-Logarithmic Quantizer for Sub-Band Coding of Audio”, *Information Technology and Control*, vol. 47 (1), pp. 131-139, 2018.

[65] Z. Perić, D. Aleksić, M. Stefanović, J. Nikolić, “New Approach to Support Region Determination of the μ-law Quantizer”, *Elektronika Ir Elektrotechnika*, vol. 19 (8), pp. 111-114, 2013.

[66] D. Aleksić, Z. Perić, J. Nikolić, “Support Region Determination of the Quasilogarithmic Quantizer for Laplacian Source”, *Przeglad Elektrotechniczny*, vol. 88(7), pp. 130-132, 2012.

[67] Z. Perić, J. Nikolić, A. Mosić, S. Panić, “A Switched -Adaptive Quantization Technique Using μ-Law Quantizers”, *Information Technology and Control*, vol. 39 (4), 317-320, 2010.

[68] M. Tančić, Z. Perić, N. Simić, S. Tomić, "Performance of Quasi-Logarithmic Quantizer for Discrete Input Signal", *Information Technology and Control,* vol. 46 (3), 395- 402, 2017.
[69] S. Tomić, Z. Perić, M. Tančić, J. Nikolić, "Backward Adaptive and Quasi-Logarithmic Quantizer for Sub-Band Coding of Audio", *Information Technology and Control,* vol. 47 (1), 131-139, 2018.
[70] Z. Perić, M. Petković, M. Dinčić, "Simple Compression Algorithm for Memoryless Laplacian Source Based on the Optimal Companding Technique", *Informatica*, vol. 20 (1), pp. 99-114, 2009.
[71] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]," *IEEE Signal Processing Magazine*, 29, 141-142, 2012.
[72] A. F. Agarap, Deep Learning Using Rectified Linear Units (ReLU). arXiv 2019, arXiv:1803.08375.
[73] Python Software Foundation. Python Language Reference, Version 2.7. Available online: http://www.python.org.(accessed on 1 September 2021).
[74] Available online: https://github.com/zalandoresearch/fashion-mnist (accessed on 10 December 2021).
[75] H. Xiao, K. Rasul, R. Vollgraf, "Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms", arXiv 2017, arXiv:1708.07747.
[76] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, C.S. Corrado, A. Davis, et al. "Tensorflow: Large-scale Machine Learning on Heterogeneous Distributed Systems", arXiv 2016, arXiv:1603.04467.
[77] Z. Perić, M. Savić, N. Simić, B. Denić, V. Despotović, "Design of a 2-bit Neural Network Quantizer for Laplacian Source", *Entropy,* vol. 23 (8), 2021.
[78] Z. Perić, B. Denić, M. Savić, V. Despotović, "Design and Analysis of Binary Scalar Quantizer of Laplacian Source with Applications", *Information,* vol. 11 (11), 501, 2020.
[79] Y. Bhalgat, J. Lee, M. Nagel et al., "LSQ+: Improving Low-bit Quantization through Learnable Offsets and Better Initialization", *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, WA, USA, 14- 19, June 2020.

[80] P. Pham, J. Abraham, J. Chung, “Training Multi-Bit Quantized and Binarized Networks with a Learnable Symmetric Quantizer,” *IEEE Access*, vol. 9, 47194-47203, 2021.

[81] D. Aleksić, Z. Perić, “One-Bit Quantizer Parametrization for Arbitrary Laplacian Sources”, Facta Universitatis Series Automatic Control and Robotics, accepted, 2022.

8. Биографија аутора

Данијела Р. Алексић је рођена у Прокупљу 3. августа 1977. године. Основну школу „9. октобар“ је завршила у Прокупљу, као ђак генерације и носилац Вукове дипломе. Гимназију у Прокупљу је такође је завршила као носилац Вукове дипломе. Као студент Електронског факултета у Нишу од 1996. године, била је добитник похвала за полагање испита у року и са високим просеком, као и добитник награде државе Норвешке за најбоље студенте 2001. године. Након одбране дипломског рада на катедри за Телекомуникације априла 2002, била је ангажована на истој катедри као истраживач стипендиста Министарства за науку и заштиту животне средине Републике Србије до септембра 2003. Тада започиње радну каријеру у компанији „Телеком Србија“, најпре у Београду, а затим у Нишу. Током рада у компанији је имала прилике да се усавршава у Мађарској, САД, Немачкој, Италији, Шведској и Републици Ирској, највише у области мобилних мрежа. Тренутно ради на позицији експерта за бежичну приступну мрежу. Удата је и мајка двоје деце.

ИЗЈАВА О АУТОРСТВУ

Изјављујем да је докторска дисертација, под називом

РАЗВОЈ КОДЕРА ТАЛАСНОГ ОБЛИКА ЗА ПОТРЕБЕ НЕУРОНСКИХ МРЕЖА И ОБРАДУ СИГНАЛА

која је одбрањена на Електронском факултету Универзитета у Нишу:

- резултат сопственог истраживачког рада;
- да ову дисертацију, ни у целини, нити у деловима, нисам пријављивала на другим факултетима, нити универзитетима;
- да нисам повредила ауторска права, нити злоупотребила интелектуалну својину других лица.

Дозвољавам да се објаве моји лични подаци, који су у вези са ауторством и добијањем академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада, и то у каталогу Библиотеке, Дигиталном репозиторијуму Универзитета у Нишу, као и у публикацијама Универзитета у Нишу.

У Нишу, _____

Потпис аутора дисертације

_____

Др Данијела Р. Алексић

ИЗЈАВА О ИСТОВЕТНОСТИ ШТАМПАНОГ И ЕЛЕКТРОНСКОГ ОБЛИКА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Наслов дисерације:

РАЗВОЈ КОДЕРА ТАЛАСНОГ ОБЛИКА ЗА ПОТРЕБЕ НЕУРОНСКИХ МРЕЖА И ОБРАДУ СИГНАЛА

Изјављујем да је електронски облик моје докторске дисертације, који сам предала за уношење у **Дигитални репозиторијум Универзитета у Нишу,** истоветан штампаном облику.

У Нишу, _____

Потпис аутора дисертације

_____

Др Данијела Р. Алексић

ИЗЈАВА О КОРИШЋЕЊУ

Овлашћујем Универзитетску библиотеку „Никола Тесла“ да у Дигитални репозиторијум Универзитета у Нишу унесе моју докторску дисертацију, под насловом:

РАЗВОЈ КОДЕРА ТАЛАСНОГ ОБЛИКА ЗА ПОТРЕБЕ НЕУРОНСКИХ МРЕЖА И ОБРАДУ СИГНАЛА

Дисертацију са свим прилозима предала сам у електронском облику, погодном за трајно архивирање.

Моју докторску дисертацију унету у Дигитални рапозиторијум Универзитета у Нишу, могу користити сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons), за коју сам се одлучила.

1. Ауторство (**CC BY**)
2. Ауторство - некомерцијално (**CC BY-NC**)
3. **Ауторство – некомерцијално – без прераде (CC BY-NC-ND)**
4. Ауторство - некомерцијално – делити под истим условима (**CC BY-NC-SA**)
5. Ауторство – без прераде (**CC BY-ND**)
6. Ауторство – делити под истим условима (**CC BY-SA**)

У Нишу, _____

Потпис аутора дисертације

_____

Др Данијела Р. Алексић