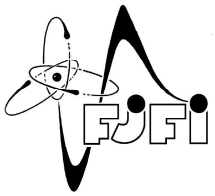
České vysoké učení technické
v Praze
Fakulta jaderná a fyzikálně inženýrská



Statistické zpracování simulovaných dat v částicové fyzice

Statistical Data Processing of Simulations in High Energy Physics

Diplomová práce

Autor: **Kristina Jarůšková**

Vedoucí práce: **Ing. Václav Kůs, Ph.D.**

Akademický rok: 2020/2021

Katedra: matematiky Akademický rok: 2020/2021

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student:	Bc. Kristina Jarůšková
Studijní program:	Aplikace přírodních věd
Studijní obor:	Aplikované matematicko-stochastické metody
Název práce (česky):	Statistické zpracování simulovaných dat v částicové fyzice
Název práce (anglicky):	Statistical Data Processing of Simulations in High Energy Physics

Pokyny pro vypracování:

1) Pokračujte ve vývoji specifické problematiky testů homogenity převážených dat v částicové fyzice (HEP) s využitím jádrových odhadů hustot (KDE) a přegenerování vzorků.
2) Proveďte srovnávací simulaci pro ověření kvality příslušných statistických algoritmů založených na odhadech vážených hustot pravděpodobností pro vybrané standardní rodiny distribucí s případnou aplikací na HEP data.
3) Proveďte rešerši technik simulací detektorů v částicové fyzice, především se zaměřte na problematiku generativních kompetitivních sítí (GANs) a jejich využití v HEP.
4) Zaměřte se na metody evaluace kvality generativních sítí v kontextu částicové fyziky a simulací snímků kalorimetrů. Prozkoumejte možnosti využití principu paradoxu narozenin pro odhad supportu generativních sítí a navrhněte modifikaci vhodnou pro simulace kalorimetru.
5) Seznamte se s modelem generativní sítě 3DGAN vyvíjeným v laboratoři CERN. Aplikujte pro tento model vybrané metody porovnání kvality simulací a srovnejte model generativní sítě s Monte Carlo přístupem.

Doporučená literatura:

1) P. Bouř, V. Kůs, Computer simulation on homogeneity testing for weighted data sets used in HEP. In 'Journal of Physics: Conference Series 1085', Bristol: IOP Publishing Ltd., 2018.

2) I. Goodfellow, Y. Bengio, A. Courville, Deep Learning. MIT Press, London, 2016.

3) S. Vallecorsa, Generative models for fast simulation. In 'Journal of Physics: Conference Series 1085', Bristol: IOP Publishing Ltd., 2018.

4) J. Goodfellow et al. Generative Adversarial Networks. ArXiv e-prints, 2014. [Online]. arXiv:1406.2661.

5) G. R. Khattak, S. Vallecorsa, F. Carminati, G. M. Khan, Particle Detector Simulation using Generative Adversarial Networks with Domain Related Constraints. In '2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)', Boca Raton, FL, USA, 2019, 28-33.

6) S. Arora, Y. Zhang, Do GANs actually learn the distribution? An empirical study. ArXiv e-prints, 2017. [Online]. arXiv:1706.08224v2.

Jméno a pracoviště vedoucího diplomové práce:

Ing. Václav Kůs, Ph.D.
Katedra matematiky, FJFI ČVUT v PRAZE, Trojanova 13, 120 00 Praha 2

Jméno a pracoviště konzultanta:

Datum zadání diplomové práce: 28.2.2021

Datum odevzdání diplomové práce: 5.1.2022

Doba platnosti zadání je dva roky od data zadání.

V Praze dne 28.2.2021

.....
garant oboru

.....
vedoucí katedry

.....
děkan

Poděkování:

Ráda bych na tomto místě poděkovala svému školiteli Ing. Václavu Kůsovi, Ph.D. za odborné vedení, nekonečnou trpělivost a velmi vstřícný přístup při vedení mé diplomové práce. Dále mé díky patří Dr. Sofii Vallecorse za mé dočasné přijetí do týmu v CERN openlab a zapojení do vývoje modelu 3DGAN.

Čestné prohlášení:

Prohlašuji, že jsem tuto práci vypracovala samostatně a že jsem uvedla všechnu použitou literaturu.

V Praze dne 3. května 2021

.....

Kristina Jarůšková

Název práce: **Statistické zpracování simulovaných dat v částicové fyzice**

Autor: Kristina Jarůšková

Obor: Aplikované matematicko-stochastické metody

Druh práce: Diplomová práce

Vedoucí práce: Ing. Václav Kůs, Ph.D., ČVUT v Praze, FJFI, Katedra matematiky

Abstrakt: Simulace elementárních částic hrají klíčovou roli při objevování nových zákonů fyziky, proto je kladen velký důraz na jejich přesnost. Standardním přístupem pro získání simulací je použití algoritmů na principu Monte Carlo (MC). MC simulace mají často podobu vážených souborů dat, což znemožňuje použití obvyklých testů homogenity pro kontrolu jejich shody s reálnými měřeními. Cílem práce bylo pomocí numerických simulací ověřit funkčnost dříve navržených modifikací testů homogenity, stejně tak jako ověření fungování nového přístupu k testování vážených dat využívajícího jádrových odhadů. Druhá část práce se soustředí na použití generativních kompetitivních sítí pro rychlé simulace kalorimetru, navíc byla implementována metoda pro validaci takto získaných simulací.

Klíčová slova: částicová fyzika, generativní kompetitivní sítě, jádrové odhady, přeuspořádání, simulace v HEP, testy homogenity, vážené vzorky dat

Title: **Statistical Data Processing of Simulations in High Energy Physics**

Author: Kristina Jarůšková

Abstract: Simulations of elementary particles play a key role in the attempts to discover new laws of physics, which is why high precision is expected of them. The common approach to generating simulations is the use of the Monte Carlo-based algorithms (MC). However, MC simulations are usually in the form of a weighted dataset making it impossible to use standard homogeneity tests to verify an agreement between simulations and real measurements. The goal of this thesis is to assemble information about possible approaches to testing homogeneity of weighted data samples, propose a new approach using kernel density estimates, and use numerical simulations to verify the applicability of both the newly proposed and previous methods of weighted homogeneity testing. This thesis also briefly discusses the use of generative adversarial networks for HEP simulations. Additionally, a general method of GAN validation was applied to a specific use case of HEP simulations.

Key words: generative adversarial networks, high energy physics, homogeneity testing, kernel density estimates, re-arranging, simulations in HEP, weighted samples

Obsah

Úvod

Objevování nových jevů na poli částicové fyziky je oblast, které se věnují mnohá mezinárodní výzkumná centra v čele s evropskou laboratoří CERN (European Organization for Nuclear Research). Ústřední strategií pro zkoumání hmoty ve vesmíru, je srovnávání teoretických predikcí s reálně naměřenými daty. Teorie je v tomto srovnání zastoupena simulacemi, tedy umělými vzorky, které odpovídají teoreticky odvozeným rozdělením.

Simulace fyzikálních jevů sehrávají v získávání nových poznatků klíčovou roli, je proto kladen velký důraz na jejich přesnost. Standardně se simulace získávají pomocí algoritmů založených na principu Monte Carlo (MC). S rostoucími požadavky na počet simulací ale začínají docházet výpočetní kapacity, proto je v posledních pár letech intenzivně zkoumána možnost využití generativních kompetitivních sítí (GANs) jako metody rychlých simulací.

Výborným nástrojem pro ověření shody MC simulací s reálnými měřeními jsou statistické testy homogenity. Problém nastane, pokud v souboru MC simulaci má každé pozorování přiřazeno také svou váhu. V tu chvíli nelze klasické testy homogenity použít. První část práce se proto věnuje problematice zobecnění testů homogenity na vážená data.

Na teoretický úvod do testů homogenity navazuje popis dosavadních přístupů k jejich zobecnění, konkrétně zobecnění neparametrických testů modifikací testovací statistiky [38] a metody transformace váženého souboru dat na nevážený pomocí tzv. přeuspořádání [14]. Navržen je také vlastní přístup k testování vážených souborů dat s využitím vážených jádrových odhadů.

Předchozí přístupy ke zobecnění testů homogenity i nově navržená varianta s jádrovými odhady jsou aplikovány na Kolmogorovův-Smirnovův test homogenity a jsou následně podrobeny numerické analýze s cílem ověřit jejich fungování pro vybrané rodiny rozdělení, nalézt výhody a limity jednotlivých variant.

Druhá část práce se zaměřuje na trochu odlišné simulace v HEP, které mají spíše podobu obrazových dat. U takového typu dat se nabízí využití GANs jako metody rychlé simulace. Práce proto obsahuje stručný úvod do fungování generativních neuronových sítí. Na ten je posléze navázáno popisem konkrétního modelu 3DGAN, který umožňuje generování simulací odezvy části detektoru navrženého pro budoucí urychlovač částic CLIC (Compact Linear Collider) v CERN.

Použití GANs pro simulace v částicové fyzice je stále poměrně novou disciplínou, chybí proto jednotný způsob evaluace takto generovaných simulací. V práci je popsán koncept metody pro evaluaci kvality sítě GAN založený na narozeninovém problému, který byl poprvé prezentován v [10]. Metoda je poté adaptována pro použití na modelu 3DGAN s cílem získat nové poznatky o jeho schopnosti generovat realistické simulace částic.

Kapitola 1

Simulace v částicové fyzice

Základním pilířem částicové fyziky je teorie nazývaná standardní model. Ten si klade za cíl popsat elementární subatomární částice tvořící veškerou hmotu ve vesmíru a vysvětlit interakce, ke kterým mezi těmito částicemi dochází. Protože se jedná pouze o teorii, je nutné jeho podobu ověřovat a podle nových experimentálních poznatků standardní model upravovat. Za tímto účelem bylo postaveno několik laboratoří a výzkumných zařízení v různých zemích, jmenujme například CERN (European Organisation for Nuclear Research) ležící na hranicích Francie a Švýcarska nebo Fermilab (Fermi National Accelerator Laboratory) a BNL (Brookhaven National Laboratory) ve Spojených státech amerických. Pomocí soustav urychlovačů částic a komplexních detektorů jsou v těchto laboratořích urychlovány a sráženy částice s cílem zjistit, zda doposud objevené částice jsou skutečně fundamentální a zda aktuální podoba standardního modelu včetně jeho parametrů souhlasí se získanými měřeními.

Srovnání naměřených dat se standardním modelem probíhá skrze tzv. *simulace*. Podoba standardního modelu a jeho parametry slouží jako vstupní informace pro algoritmy, pomocí kterých lze simulovat interakce částic a jejich chování v detektorech různého druhu. V komparativních analýzách pak tyto simulace zastupují zkoumanou teorii standardního modelu a jsou porovnávány s výsledky reálných měření. Pokud je mezi simulovanými a skutečnými daty nalezena dostatečná shoda, pak teorie nebyla vyvrácena a považujeme ji za platnou. Pokud mezi simulacemi a reálnými daty pozorujeme drobné, ale nezanedbatelné odchylky, obvykle dochází k upřesnění parametrů standardního modelu. Simulovaná a měřená data mohou být také ve výraznějším rozporu, což může ukazovat na chybu v samotné konstrukci modelu a může to znamenat objevení nové částice nebo síly [17]. Ověření správnosti modelu ale není jediným účelem generování simulací. Simulace se využívají také při návrhu detektorů, vývoji softwarových nástrojů, které surová data z detektorů zpracovávají, a v neposlední řadě také při návrhu a ladění kroků fyzikální analýzy naměřených dat [40]. Kvalitní simulace odpovídající aktuální podobě standardního modelu jsou tedy nezbytnou součástí celého procesu získávání nových poznatků v oblasti částicové fyziky.

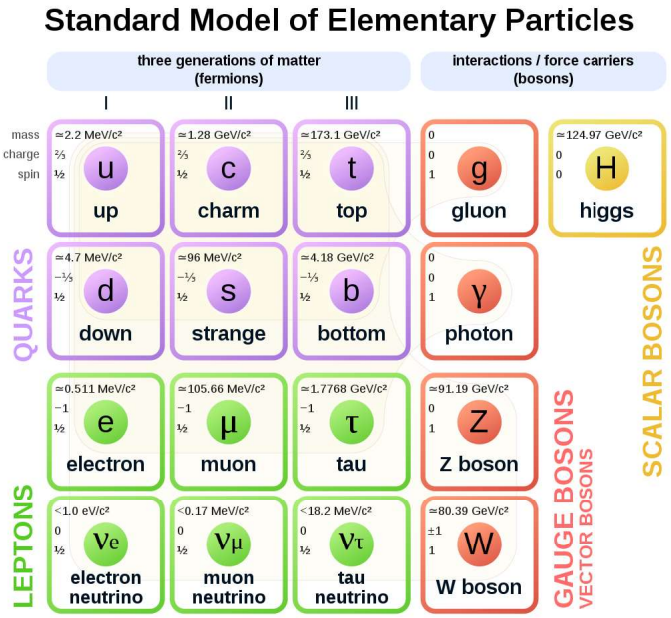
Klasickým přístupem ke generování simulovaných dat jsou tzv. Monte Carlo algoritmy. Tento přístup využívá podoby standardního modelu a jeho parametrů a poskytuje simulace srážek částic a jejich chování při průchodu detektorem. Výhodou je velmi dobrá přesnost, ale s rostoucím počtem potřebných simulací se čím dál tím větší překážkou stává značná časová náročnost tohoto přístupu (kompletní simulace jedné interakce trvá cca 1 minutu [20]). V posledních letech jsou proto hledány nejen způsoby, jak výpočty

založené na Monte Carlo algoritmech zrychlit, ale také alternativní metody generování simulovaných dat. Oblastí velkého zájmu v HEP komunitě se stávají generativní modely spadající do kategorie hlubokého učení (deep learning), například variační autoencodery (variational autoencoders) nebo generativní kompetitivní sítě (generative adversarial networks, zkráceně GANs) [4].

V následujícím textu první kapitoly je nejprve krátce představena současná podoba standardního modelu částicové fyziky. Dále jsou poskytnuty základní informace o Monte Carlo simulacích. Poslední část této kapitoly shrnuje poznatky o metodách rychlé simulace v HEP s využitím generativních algoritmů.

1.1 Standardní model částicové fyziky

Současná verze standardního modelu na obrázku 1.1 obsahuje fundamentální částice a interakce, které byly doposud experimentálně pozorovány. Tyto fundamentální částice dělíme do dvou skupin - kvarky (quarks) a leptony (leptons). Každá z těchto skupin obsahuje 6 částic, které jsou sdružené po dvojicích do tzv. generací. První generace částic je tvořena nejlehčími a nejstabilnějšími částicemi a tvoří veškerou stabilní hmotu, naopak druhá a třetí generace obsahují těžší a méně stabilní částice, které se zpravidla rychle rozpadají na částice stabilnější. Nejstabilnějšími z kvarků jsou up kvark a down kvark, následuje dvojice charm a strange kvark, nejméně stabilní jsou pak top kvark a bottom kvark. Do třech generací dělíme rovněž leptony - nejstabilnější jsou elektron a elektronové neutrino, dále mion a mionové neutrino a tauon a tauonové neutrino. Kvarky a leptony dohromady nazýváme také jako fermiony.



Obrázek 1.1: Standardní model částicové fyziky [42]. Z tabulky je patrné rozdělení částic na femiony (fermions) a bosony (bosons). Fermiony jsou dále rozděleny na kvarky (quarks) a leptony (leptons).

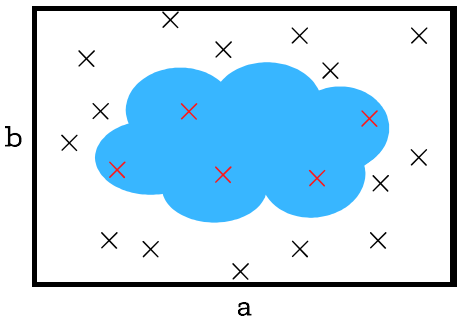
Veškeré interakce, které probíhají mezi výše uvedenými částicemi, lze vysvětlit pomocí čtyř základních sil - silná interakce, slabá interakce, elektromagnetická síla a gravitační síla. Silná a slabá interakce působí pouze na subatomární úrovní, naproti tomu elektromagnetická a gravitační síla mají neomezený dosah působení. Pokud bychom měli tyto interakce seřadit vzestupně dle jejich síly, pak gravitace je nejslabší, následuje slabá interakce, elektromagnetická síla a jak již název napovídá, nejsilnější je silná interakce. Ke každému typu interakce jsou přiřazeny částice, které fungují jako její zprostředkovatelé. Tyto částice nesou souhrnné označení bosony. Silnou interakci zprostředkovává částice nazývaná gluon, za nositele elektromagnetické interakce je považován foton a bosony W a Z zprostředkovávají slabou interakci. Předpokládá se, že zprostředkovatelem gravitace je částice nazvaná graviton, existenci tohoto bosonu se ale zatím nepodařilo ověřit. Nejmladší objevenou částicí je Higgsův boson, jehož nalezení bylo oficiálně potvrzeno 4. července 2012 [19] a již v roce 2013 byla za předpovězení jeho existence udělena Nobelova cena dvojici teoretických fyziků Peteru Higgsovi a Françoisovi Englertovi.

Standardní model v současné podobě neposkytuje kompletní popis chování hmoty ve vesmíru. Jak již bylo naznačeno výše, do teorie standardního modelu se zatím nepodařilo zahrnout působení gravitační síly a popsat tak všechny interakce jednotnou formulací. Model tedy velmi dobře funguje na úrovni mikrosvěta, ve kterém je vliv gravitační síly zanedbatelný, ale u větších těles se gravitační síla projevuje více a zákonitosti vyplývající ze standardního modelu nelze použít. Stejně tak standardní model nevysvětluje existenci temné hmoty a temné energie nebo nerovnováhu mezi hmotou a antihmotou.

1.2 Monte Carlo simulace

Jak bylo zmíněno v úvodu kapitoly, standardním metodou pro získání simulací srážek částic nebo detektorů jsou programy založené na principu Monte Carlo algoritmů. Tento přístup předpokládá, že zkoumaný systém a jeho vývoj lze popsat pomocí pravděpodobností. V principu pak namísto sestavování komplikovaných rovnic a provádění složitých výpočtů provedeme raději velké množství opakování zkoumaného experimentu za využití generátorů náhodných hodnot a z výsledků těchto opakování vyvodíme hledaný závěr.

Velmi jednoduchou ukázkou je úloha výpočtu plochy dvourozměrného obrazce znázorněná na obrázku 1.2. Pokud bychom znali rovnice popisující zadaný obrazec, lze jeho obsah zjistit pomocí dvourozměrné integrace. Druhou možností je opakované generování souřadnic náhodných bodů v rámci vyobrazeného obdélníku o obsahu ab z uniformního rozdělení s následnou kontrolou, zda daný bod leží uvnitř modrého obrazce či nikoliv. Přibližnou plochu obrazce pak získáme jako $\frac{\#\,zásahů}{\#\,pokusů}\,ab$. Monte Carlo simulace v částicové fyzice fungují na obdobném principu, ale simulované systémy a výpočty jsou výrazně komplikovanější a pro tyto účely jsou vyvíjeny samostatné softwarové nástroje.



Obrázek 1.2: Ilustrace výpočtu plochy obrazce generováním náhodných bodů.

Využití Monte Carlo simulací v HEP lze rozdělit na dvě základní skupiny. První z nich je simulování fundamentálních interakcí částic, především srážek částic nebo jejich rozpadů. Pomocí těchto simulací jsou vypočítávány například účinné průřezy (neboli pravděpodobnosti, že ostřelující částice bude jistým způsobem interagovat s částicí terče) nebo takzvané vektory čtyřhybností pro produkty srážek a rozpadů. Pro simulace těchto fundamentálních interakcí jsou vyvíjeny různé softwarové nástroje, jmenujme například PYTHIA 8 [35], Herwig 7 [12] nebo program Sherpa [23].

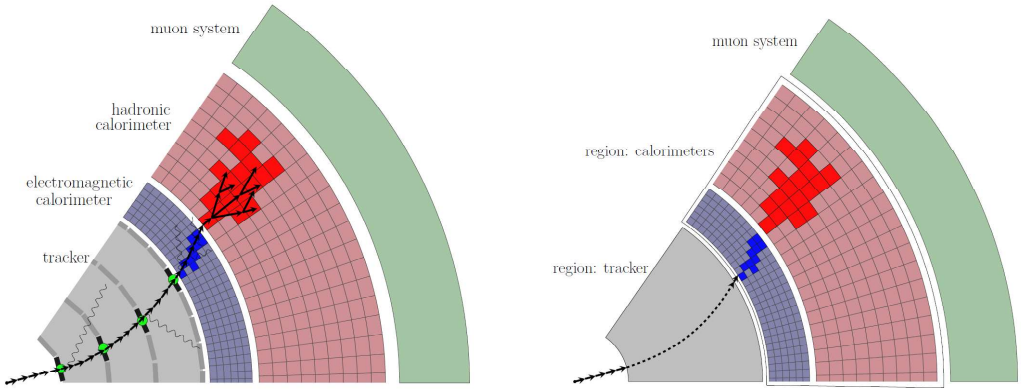
Simulace fundamentálních interakcí slouží jako vstupní údaje pro druhou skupinu simulací, kterými jsou simulace detektorů. Jedná se o modelování průchodu částice detektorem včetně interakcí této částice s materiálem detektoru, vzniku sekundárních částic a jejich cesty skrze detektor, a to až do momentu rozpadu, pohlcení nebo opuštění detektoru původní částicí i všemi sekundárními. Celý průběh simulace závisí nejen na vlastnostech primární částice, ale také na geometrii a materiálu příslušného detektoru. Nejčastěji používaným nástrojem pro tento typ simulací je program Geant4 [6], aktuálně je ve vývoji i novější verze GeantV [7].

1.2.1 Úplné a rychlé simulace

Metody simulace detektorů je možné dále rozdělit na dva druhy. Takzvaná **úplná simulace** znamená, že na počátku máme částici vstupující do kalorimetru s přesně určenými počátečními vlastnostmi a po krátkých časových úsecích modelujeme, jak se mění poloha částice, její hybnost a jaká je odezva detektoru. Zároveň je v každém kroku simulováno, zda dojde k interakci částice s materiálem detektoru, jejímu rozpadu nebo zda částice pokračuje dál v původním směru. Pokud dojde k interakci částice s materiálem kalorimetru nebo k jejímu rozpadu za vzniku sekundárních částic, jsou modelovány hybnosti, souřadnice a odezva detektoru i pro všechny sekundární částice. Některé částice mohou vyprodukovat miliony sekundárních částic a při úplné simulaci je pro všechny stejně detailně modelován jejich průchod detektorem. Jedná se tedy o velmi podrobnou simulaci, která je časově i výpočetně velmi náročná.

Náročné úplné simulace nejsou vždy nezbytně nutné, například pro předběžné analýzy jsou často využívány tzv. **rychlé simulace**. Jedná se o simulace, které jsou méně detailní než simulace úplné, proto je jejich výpočet rychlejší. Jednou z možností, jak dosáhnout zrychlení simulování na úkor přesnosti, je místo modelování průchodu částice krok po kroku včetně jejích hybností a dalších charakteristik simulovat rovnou pouze odpověď detektoru na částici a produkty jejího rozpadu jako celek. Na obrázku 1.3 je ilustrován rozdíl mezi úplnou a rychlou simulací. První vnitřní vrstvou vyobrazeného detektoru je tzv. tracker, který snímá částice s elektrickým nábojem, modře a červeně jsou znázorněny dva různé druhy kalorimetrů (elektromagnetický a hadronový), které zachycují energii vyzářenou částicemi. Zelenou barvou je znázorněna vrstva detektoru, která slouží především k zachycení mionů. Pro úplnou simulaci jsou zaznamenány hybnosti a energie pro prvotní částici i všechny sekundární částice stejně jako energie zaznamenané detektorem, v rychlé simulaci byla modelována pouze odezva detektorů.

Další možností, jak dosáhnout zkrácení výpočetního času při modelování detektoru, je použití zjednodušeného popisu geometrie detektoru nebo nasimulování produktů častých rozpadů předem a jejich použití v případě, že při simulaci krok po kroku k danému rozpadu dojde. Aktuálně používaný softwarový nástroj Geant4 již tyto možnosti zrychlení simu-



(a) Úplná simulace

(b) Rychlá simulace

Obrázek 1.3: Ilustrace rozdílu mezi úplnou a rychlou simulací detektoru [43]. Při úplné simulaci (a) byl průchod částice modelován krok po kroku včetně odezvy detektoru a hybností všech částic (primárních i sekundárních), u rychlé simulace (b) byla modelována pouze odezva detektoru.

lací nabízí. Velká pozornost je také věnována optimalizaci všech výpočtů při simulacích a efektivnímu využití aktuálního hardwaru.

1.3 Generativní algoritmy

Monte Carlo nástroje popsané v předchozí podkapitole představují metody simulace dosahující vysoké přesnosti, ale jejich výpočet je velmi časově náročný. Současné výpočetní kapacity by při používání stávajících metod byly brzy nedostatečné a možnosti rychlých simulací metodou Monte Carlo nenabízejí dostatečné zrychlení výpočtů. Proto se do centra pozornosti dostaly generativní algoritmy založené na neuronových sítích, které by mohly poskytnout dostatečné zrychlení produkce simulací.

1.3.1 Motivace

Evropská laboratoř CERN má pro potřeby analýz a simulací k dispozici tzv. Worldwide LHC Computing Grid (WLCG), neboli hardwarovou a softwarovou infrastrukturu, která umožňuje distribuovat výpočty na individuální stroje umístěné ve 170 různých lokacích ve 42 zemích světa. Přibližně 50 % výpočetní kapacity WLCG je v současnosti využíváno pro simulace částic a detektorů, zbylá kapacita je využívána pro analýzy měřených dat [40]. Na rok 2027 je plánováno spuštění velkého hadronového urychlovače (Large Hadron Collider, zkráceně LHC) po provedení série úprav, díky kterým bude mít 10× vyšší luminositu než v prvním roce svého fungování [18]. Předpokládá se, že na tzv. HL-LHC (High Luminosity LHC) bude možné získat 150× více dat ve srovnání s původní konstrukcí LHC. Z tohoto důvodu porostou požadavky na počet potřebných simulací, ale s výrazným rozšiřování výpočetních kapacit WLCG se naopak nepočítá [5].

Předpokládaný nedostatek výpočetních kapacit v budoucnu je hlavní motivací pro optimalizaci rychlých simulací a hledání nových přístupů. Možnosti optimalizací rychlých

simulací popsané v předchozí podkapitole a implementované v programu GEANT4 nabízí zrychlení různých řádů (10× − 1000×), ale obvykle jsou aplikovatelné pouze na velmi specifické úlohy v rámci konkrétního experimentu, chybí jednotný přístup produkce rychlých simulací. V posledních letech proto probíhá výzkum využitelnosti generativních algoritmů založených na hlubokém učení a neuronových sítích pro simulace detektorů. První výsledky naznačují, že při rychlých simulacích elektromagnetických kalorimetrů pomocí deep learningu je možné dosáhnout zrychlení produkce simulací až o 6 řádů, zatím stále za cenu ztráty přesnosti [40].

Použití generativních algoritmů pro vytváření simulací je většinou zkoumáno na úloze simulace elektromagnetického kalorimetru (EM kalorimetru), protože simulace této části detektoru je výpočetně nejnáročnější (při simulacích pro detektor ATLAS v CERN je 70 % času věnováno EM kalorimetru [43]).

1.3.2 Aplikace GAN v HEP

Elektromagnetický kalorimetr je zařízení, který slouží především k zachycení elektronů a fotonů a zaznamenání jejich energie vyzářené při průchodu materiálem kalorimetru. Zároveň je možné energie měřené v různých bodech kalorimetru interpretovat jako intenzitu pixelů trojrozměrného šedotónového obrázu. Právě tato analogie mezi snímky kalorimetru a obrazovými daty vybízí k využití generativních modelů používaných při zpracování obrazu, konkrétně *generativních kompetitivních sítí* (generative adversarial networks, zkráceně GANs) [25].

Základní strukturu GAN tvoří trénovací data a dvě neuronové sítě: diskriminační síť (anglicky discriminator) a generátor (anglicky generator). Cílem generátoru je naučit se rozdělení trénovacích dat a produkovat nové obrázky ze stejného rozdělení, naopak diskriminační síť se snaží trénovací data a generované obrázky od sebe rozlišit, čímž poskytuje zpětnou vazbu na kvalitu generátoru. V principu se tedy jedná o úlohu minimax optimalizace, která v ideálním případě končí nalezením optimálních parametrů sítí, pro které je nastavením generátoru dosaženo minimalizace maximální rozlišovací schopnosti diskriminační sítě. Podrobnějšímu popisu fungování GAN je věnována samostatná kapitola 4.

Současné modely GAN dokážou úspěšně generovat kvalitní obrázky lidských tváří, zvířat či objektů, které jsou velmi blízko skutečným fotografiím. V porovnání s těmito aplikacemi má simulování EM kalorimetrů určitá specifika. Zaprvé, intenzity pixelů pro odezvu kalorimetru (tedy snímané hodnoty energií) nabývají hodnot s rozsahem několika řádů. Zadruhé, i když simulovaná sprška obsahuje miliony částic, tak výskyt pixelů s nenulovou hodnotou intenzity je v celkovém obrazu kalorimetru řídký.

Model LAGAN

Jedním z prvních úspěšných příkladů použití GAN pro simulace kalorimetru je model LAGAN (Location-Aware Generative Adversarial Network [20]). Pomocí tohoto modelu se podařilo simulovat 2D snímky sprškek částic v kalorimetru, přičemž simulované hodnoty intenzit jednotlivých pixelů kopírovaly rozsah pozorovaný v trénovacích datech, který byl přibližně 5 řádů ($\sim 10^{-3}$ až 10^{2} GeV). Byla pozorována také dobrá shoda mezi generovanými a trénovacími daty z pohledu četnosti hodnot v konkrétních intervalech.

Model CaloGAN

Za další úspěšnou aplikaci lze považovat model CaloGAN [32], pomocí kterého byly simulovány spršky částic pro zjednodušenou geometrii EM kalorimetru detektoru ATLAS. Daný typ kalorimetru je sestaven ze třech vrstev, pro simulaci odezvy každé z nich byl proto využit předchozí model LAGAN doplněný o přenosovou jednotku pro zachování korelace mezi jednotlivými vrstvami. V modelu CaloGAN bylo zároveň testováno podmínění generovaného obrázku zadáním počáteční energie vstupní částice a bylo pozorováno, že pomocí CaloGAN lze generovat věrohodné simulace také pro částice s počáteční energií mimo původní rozsah hodnot použitý v trénovacích datech. Toto pozorování naznačovalo, že podmíněný generativní model je schopný určité extrapolace vzhledem k vstupnímu parametru počáteční energie. Přestože trénování modelů hlubokého učení je časově náročné, tak samotné generování nových snímků pomocí CaloGAN je o 5 řádů rychlejší v porovnání s odpovídající Monte Carlo simulací.

Model 3DGAN

V rámci skupiny CERN openlab je aktuálně vyvíjen model s názvem 3DGAN [16], pomocí kterého lze přímo generovat 3D snímky kalorimetru, který bude instalován na budoucím lineárním urychlovači částic CLIC (Compact Linear Collider). Možnost podmínění generovaného obrázku počáteční energií vstupní částice je pro tento model rozšířena o podmínění úhlem, pod kterým původní částice do kalorimetru vletí vzhledem k jeho čelní stěně. Tento model představuje první koncept metody rychlé simulace založené na GAN, která v budoucnu nabídne uživateli simulačního balíku možnost přizpůsobení parametrů detektoru i částice. Podrobnějšímu popisu modelu 3DGAN je věnována samostatná podkapitola 4.3.

První výsledky použití GAN pro simulace EM kalorimetrů, které tvoří časově nejnáročnější část simulací, jsou velmi slibné [4]. Generované snímky vykazují v mnoha aspektech dobrou shodu s Monte Carlo simulacemi a podle předběžných analýz nabízí zrychlení simulačního procesu až o 5 řádů.

Otevřeným problémem zůstává zvolení jednotného přístupu k vyhodnocování kvality generovaných obrázků. Vzhledem k povaze dat, tedy 3D snímky s velkým rozptylem intenzit pixelů a řídkosti nenulových hodnot, není jednoznačné, jak kvalitu snímků přímo porovnat. Pro většinu analýz bylo doposud použito především nepřímé porovnání skrze rekonstruování fyzikálních veličin (například celkové hmotnosti interagujících částic nebo celkové energie) a srovnání jejich distribuce s hodnotami spočtenými na trénovacích datech. Generativním kompetitivním sítím a validaci jejich výsledků v kontextu HEP je věnována kapitola 4.

Kapitola 2

Testy homogenity

Předchozí kapitola shrnula základní informace o simulacích v HEP a zdůraznila jejich důležitost při objevování nových zákonitostí chování hmoty. Pokud se zaměříme na simulace detektorů, tak formát těchto dat je stejný, jako výstupy skutečných měření. V případě kalorimetru jde tedy o hodnotu signálu z čidel, která zaznamenávají energii vyzářenou částicemi. Kombinací těchto elektrických signálů z detektorů, znalosti jejich konstrukce a informací o vstupních parametrech experimentu lze následně rekonstruovat různé veličiny, například úhly mezi dráhami částic nebo hybnosti. Rekonstruované fyzikální veličiny se posléze používají jako vstupy do dalších kroků zpracování dat.

Monte Carlo simulace detektorů a z nich zrekonstruované hodnoty fyzikálních veličin slouží nejen pro přímé porovnání teorie s naměřenými daty, ale také jako trénovací data pro klasifikační algoritmy různého druhu. Aby mohl být takový klasifikátor dobře natrénován, je nutné ověřit, zda simulace odpovídají reálným měřením z pohledu rekonstruovaných fyzikálních veličin, tedy otestovat shodnost rozdělení simulací a měření pro každou veličinu. Jedním z nástrojů, který lze použít, jsou *testy homogenity*.

Pokud dojde k situaci, že hodnoty fyzikálních veličin simulovaných dat neodpovídají skutečným měřením, nelze takové simulace použít k trénování klasifikačního či jiného algoritmu. K takové situaci může dojít například z důvodu trochu odlišného finálního nastavení detektoru, než pro které se původní simulace produkovaly, nebo třeba drobnou závadou detektoru, která se objeví až v průběhu měření a způsobí vychýlení snímaných hodnot. Prvním možným řešením je generovat nové simulace, které budou odpovídat novému nastavení nebo chybě detektoru, což je ale časově i výpočetně velmi náročné. Častěji se proto přistupuje k druhé možnosti, kterou je přiřazení vah jednotlivým simulovaným pozorováním tak, aby rozdělení rekonstruovaných veličin odpovídalo skutečným měřením. Po převážení simulací je shodnost rozdělení potřeba opět zkontrolovat.

Různé druhy testů homogenity jsou velmi dobrým nástrojem pro učinění rozhodnutí, zda dva soubory dat pochází ze stejného rozdělení, protože poskytují také odhad pravděpodobnosti, že učiněné rozhodnutí je správné. Jejich nevýhodou z hlediska HEP je, že ve své standardní podobě jsou aplikovatelné pouze na nevážená data. V kontextu HEP se tento problém obvykle řeší rozdělením pozorování do několika binů a testováním shodnosti rozdělení, která jsou reprezentována histogramovými odhady. Při takovém postupu ale dochází ke ztrátě informace o rozložení dat uvnitř jednotlivých binů.

V této kapitole jsou popsány tři přístupy, jak lze princip testů homogenity využít také pro data s váhami. První přístup je založený na modifikaci testovací statistiky vybraného

testu pro vážená data užitím vážené empirické distribuční funkce navržené v [37, 38]. Principem druhého přístupu je transformace váženého souboru dat na soubor neválený metodou re-arranging [13, 14] nebo pomocí jádrového odhadu (kernel density estimate, zkráceně KDE).

2.1 Klasické testy homogenity

Testy homogenity jsou speciálním případem testů hypotéz, proto nejprve popíšeme jejich obecný princip. Uvažujme pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathrm{P})$, kde Ω je neprázdná množina, $\mathcal{A}$ je σ-algebra nad Ω a P pravděpodobnostní míra definovaná na $\mathcal{A}$. Dále uvažujme náhodný výběr $X_1, X_2, \ldots, X_n$ z rozdělení F, tedy $X_1, X_2, \ldots, X_n$ jsou nezávislé a stejně rozdělené náhodné veličiny (zkráceně *i.i.d.*) $X_j : \Omega \to \mathbb{R}\ \forall j \in \{1, \ldots, n\}$, $n \in \mathbb{N}$. Testy hypotéz slouží k ověření platnosti předem stanovené nulové hypotézy H_0 proti alternativní hypotéze H_1 na základě dostupných dat, tedy konkrétní realizace $x_1, x_2, \ldots, x_n$ náhodného výběru.

V případě testů homogenity pracujeme ne s jedním, ale se dvěma náhodnými výběry, prvním $X_1, X_2, \ldots, X_n$ z rozdělení F a druhým $Y_1, Y_2, \ldots, Y_m$ z rozdělení G. Pro ověření shodnosti rozdělení tedy formulujeme hypotézy H_0 a H_1 následujícím způsobem:

$$H_0 : F = G \quad \text{vs.} \quad H_1 : F \neq G. \tag{2.1}$$

V dalším textu budeme značit $\boldsymbol{X}$ náhodný výběr $(X_1, X_2, \ldots, X_n)$, dále pomocí $\boldsymbol{x}$ označme konkrétní realizaci $(x_1, x_2, \ldots, x_n)$ náhodného výběru.

Na základě konkrétní realizace náhodného výběru, rozhodujeme o zamítnutí nebo nezamítnutí platnosti nulové hypotézy H_0. Množinu bodů $W \subset \mathbb{R}^n$, pro které zamítáme H_0, nazýváme kritickým oborem[1]. Její doplněk, tedy množinu, pro kterou H_0 nezamítáme, značíme standardně W^C. Rozhodovací pravidlo testu lze pomocí množiny W zapsat následovně:

$$\begin{aligned} \boldsymbol{x} \in W \quad &\ldots \quad \text{zamítáme } H_0, \\ \boldsymbol{x} \notin W \quad &\ldots \quad \text{nezamítáme } H_0. \end{aligned}$$

Definování kritické oblasti W je tedy klíčové pro rozhodnutí testu. Pro konstrukci rozhodovacího pravidla se v praxi obvykle nepoužívá přímo n-rozměrná kritická oblast $W \subset \mathbb{R}^n$, ale její obraz v $\mathbb{R}$ získaný zobrazením $T : \mathbb{R}^n \to \mathbb{R}$ nazývaným *testovací statistika*. Při značení $\widetilde{W} = T(W) \subset \mathbb{R}$ zamítáme H_0, pokud $T(\boldsymbol{x}) \in \widetilde{W}$.

Při rozhodování o zamítnutí hypotézy H_0 se můžeme dopustit dvou druhů chyb, které jsou znázorněné v tabulce 2.1. Chybu I. druhu, neboli zamítnutí hypotézy H_0, přestože tato platí, obvykle považujeme za závažnější, proto je kritická oblast $\widetilde{W}$ volena tak, aby platilo

$$\mathrm{P}\,(\text{chyba I. druhu}) = \mathrm{P}\left(T(\boldsymbol{X}) \in \widetilde{W} | H_0\right) = \alpha, \tag{2.2}$$

kde $\alpha \in (0, 1)$ je předem určená *hladina významnosti* testu. Definice kritické oblasti $\widetilde{W}$, respektive W, tedy závisí na volbě hladiny testu α. Pokud nelze zvolit $\widetilde{W}$ tak, abychom dosáhli rovnosti ve výrazu (2.2), volíme $\widetilde{W}$ tak, aby P (chyba I. druhu) byla co největší

[1] Pro dva náhodné výběry uvažujeme $W \subset \mathbb{R}^{n+m}$.

a zároveň stále menší než α. Při pevně stanovené hladině významnosti se následně při volbě $\widetilde{W}$ snažíme minimalizovat

$$\mathrm{P}\left(\text{chyba II. druhu}\right) = \mathrm{P}\left(T(\boldsymbol{X}) \notin \widetilde{W} | H_1\right) = 1 - \beta, \tag{2.3}$$

kde $\beta = \mathrm{P}\left(T(\boldsymbol{X}) \in \widetilde{W} | H_1\right)$ nazýváme *silou testu*.

Tabulka 2.1: Znázornění vztahu mezi rozhodnutími testu a možnými chybami.

	Nezamítáme H_0	Zamítáme H_0
H_0 platí	správně	chyba I. druhu
H_1 paltí	chyba II. druhu	správně

Rozhodování při testování je obvykle založené na hodnotě speciální charakteristiky, tzv. p-hodnotě (angl. p-value). P-hodnotu lze definovat jako infimum všech α hladin významnosti testu, na kterých ještě pro daný soubor dat $\boldsymbol{x}$ zamítáme nulovou hypotézu H_0. P-hodnotu lze také interpretovat jako pravděpodobnost, že při dalším opakování stejného měření získáme data, která budou odporovat H_0 více než testovaný soubor $\boldsymbol{x}$. Platí tedy, že pokud je pro naměřená data p-hodnota $< \alpha$, pak H_0 zamítáme, pokud je p-hodnota naopak vyšší, tak H_0 nezamítáme.

Příklad definice a použití p-hodnoty budeme ilustrovat na příkladu tzv. pravostranného testu, pro který je kritická oblast definována $\widetilde{W} = \{\boldsymbol{x} \in \mathbb{R}^n | T(\boldsymbol{x}) > C_\alpha\}$, kde C_α je hranice kritické oblasti pro test na hladině významnosti α s testovací statistikou T. Pro takový test spočteme p-hodnotu

$$p\text{-hodnota} = \mathrm{P}\left(T(\boldsymbol{X}) > T(\boldsymbol{x})\right) = 1 - F_T\left(T(\boldsymbol{x})-\right). \tag{2.4}$$

2.1.1 Neparametrické testy homogenity

Nyní se zaměříme na neparametrické testy homogenity, protože se při budoucí aplikaci chceme vyhnout nutnosti odhadovat parametrickou rodinu distribucí. Neparametrické testy pracují s empirickým odhadem kumulativní distribuční funkce, který definujeme pouze s použitím naměřených hodnot.

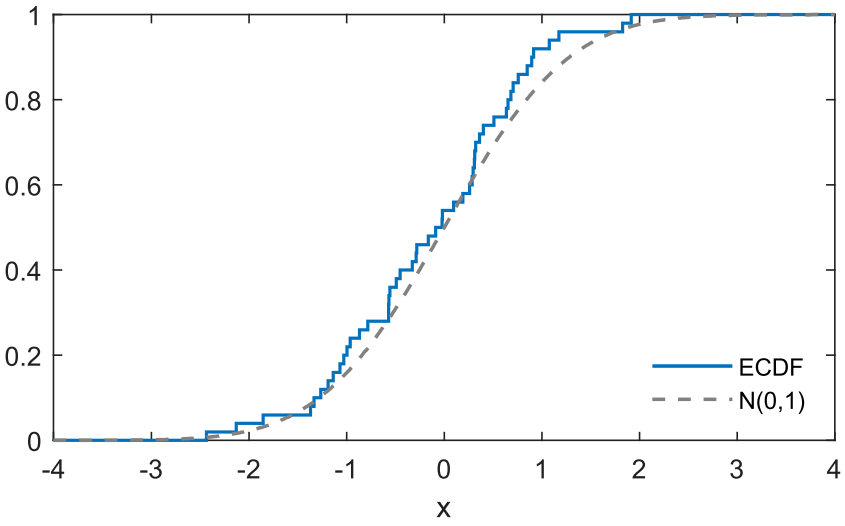
Definice 2.1.1. Nechť $X_1, X_2, \ldots, X_n$ je *i.i.d.* reálný náhodný výběr z rozdělení F. *Empirickou distribuční funkcí* (empirical cumulative distribution function, zkráceně ECDF) pro výběr $X_1, X_2, \ldots, X_n$ rozumíme

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}_{(-\infty,x]}(X_j), \quad \forall x \in \mathbb{R}, \tag{2.5}$$

kde $\mathbf{I}_{(-\infty,x]}(\cdot)$ je indikátor množiny $(-\infty, x]$.

Empirická distribuční funkce je podle definice neklesající, zprava spojitou, schodovitou funkcí, jejíž hodnota se v každém bodě X_i zvýší o $\frac{1}{n}$, příklad ECDF pro *i.i.d.* náhodný výběr z normálního rozdělení je na obrázku 2.1. Ze zákona velkých čísel pro ECDF náhodného výběru $X_1, X_2, \ldots, X_n$ z rozdělení F plyne, že

$$F_n(x) \xrightarrow{s.j.} F(x) \quad \forall x \in \mathbb{R}, \tag{2.6}$$



Obrázek 2.1: Graf empirické distribuční funkce pro 50 pozorování z rozdělení $\mathrm{N}(0,1)$.

tedy že F_n je konzistentním odhadem skutečné distribuční funkce F. Lze také dokázat, že F_n je zároveň nestranným odhadem [39].

Empirická distribuční funkce F_n konverguje ke skutečné distribuci F dokonce stejnoměrně, nejprve ale potřebujeme definovat Kolmogorovu vzdálenost dvou distribučních funkcí.

Definice 2.1.2. Nechť F a G jsou dvě distribuční funkce na $\mathbb{R}$. Pak definujeme *Kolmogorovu vzdálenost*

$$K(F,G) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|. \tag{2.7}$$

Dle Glivenkova-Cantelliho teorému pro ECDF F_n náhodného výběru z F pro $n \to +\infty$ platí

$$K(F_n, F) \xrightarrow{s.j.} 0, \tag{2.8}$$

tedy F_n stejnoměrně konverguje k F, a to pro libovolnou distribuční funkci F. Důkaz tohoto tvrzení je uveden například v [39].

Příkladem neparametrického testu homogenity, který je založený čistě na empirických distribučních funkcích, je *dvouvýběrový Kolmogorovův-Smirnovův test* (angl. two-sample Kolmogorov-Smirnov test, zkráceně KS test). Uvažujme dva náhodné výběry, $X_1, X_2, \ldots, X_n$ z distribuce F a $Y_1, Y_2, \ldots, Y_m$ z distribuce G, které jsou nezávislé. KS test slouží k ověření hypotézy

$$H_0 : (\forall x \in \mathbb{R})(F(x) = G(x)) \quad \text{vs.} \quad H_1 : (\exists x \in \mathbb{R})(F(x) \neq G(x)). \tag{2.9}$$

Testovací statistika KS testu vychází z Kolmogorovy vzdálenosti

$$K_{n,m} = \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)|, \tag{2.10}$$

kde F_n je ECDF náhodného výběru $\boldsymbol{X}$ a G_m je ECDF náhodného výběru $\boldsymbol{Y}$.

Za platnosti nulové hypotézy, tedy $F = G$, a za předpokladu spojitosti skutečného rozdělení pro testovací statistiku $K_{n,m}$ dle [29] platí

$$\sqrt{\frac{nm}{n+m}} K_{n,m} \xrightarrow[n,m\to+\infty]{\mathcal{D}} Z, \tag{2.11}$$

kde Z je náhodná veličina s distribuční funkcí

$$H(\lambda) = \begin{cases} 1 - 2\sum_{k=1}^{+\infty} (-1)^{k-1} e^{-2k^2\lambda^2} & \text{pro } \lambda > 0, \\ 0 & \text{pro } y \leq 0. \end{cases} \tag{2.12}$$

Pro Kolmogorovu vzdálenost $K(F, G)$ platí, že pro shodné distribuční funkce je rovna nule a naopak pro rozdílné distribuční funkce její hodnota roste. Hypotézu H_0 budeme zamítat, pokud je hodnota testovací statistiky vyšší než kritická hodnota $h_{1-\alpha}$ (F_n a G_m se od sebe příliš liší). Proto formulujeme tzv. pravostranné rozhodovací pravidlo

$$H_0 \text{ zamítáme} \quad \Leftrightarrow \quad \sqrt{\frac{nm}{n+m}} K_{n,m} \geq h_{1-\alpha}, \tag{2.13}$$

kde jsme použili značení $h_{1-\alpha}$ pro $(1-\alpha)$-kvantil rozdělení H. Protože známe pouze asymptotické rozdělení testovací statistiky, tak test má tzv. asymptotickou hladinu α. Příslušnou p-hodnotu, podle které se v praxi při vyhodnocování testu řídíme, spočteme jako

$$p\text{-hodnota} = 1 - H\left(\sqrt{\frac{nm}{n+m}} K_{n,m}\right). \tag{2.14}$$

Všimněme si, že pokud hypotéza H_0 neplatí, tak pro $n, m \to +\infty$

$$K_{n,m} \xrightarrow{\mathrm{P}} \sup_{x\in\mathbb{R}} |F(x) - G(x)| > 0 \Longrightarrow \sqrt{\frac{nm}{n+m}} K_{n,m} \xrightarrow{\mathrm{P}} +\infty. \tag{2.15}$$

Test zároveň nijak nezávisí na skutečných distribucích F a G a asymptoticky je síla testu rovna 1 [30].

Dalšími příklady neparametrických testů homogenity jsou Andersonův-Darlingův test a Cramer-von Misesův test, které stejně jako KS test využívají empirické distribuční funkce v definici testovací statistiky [21].

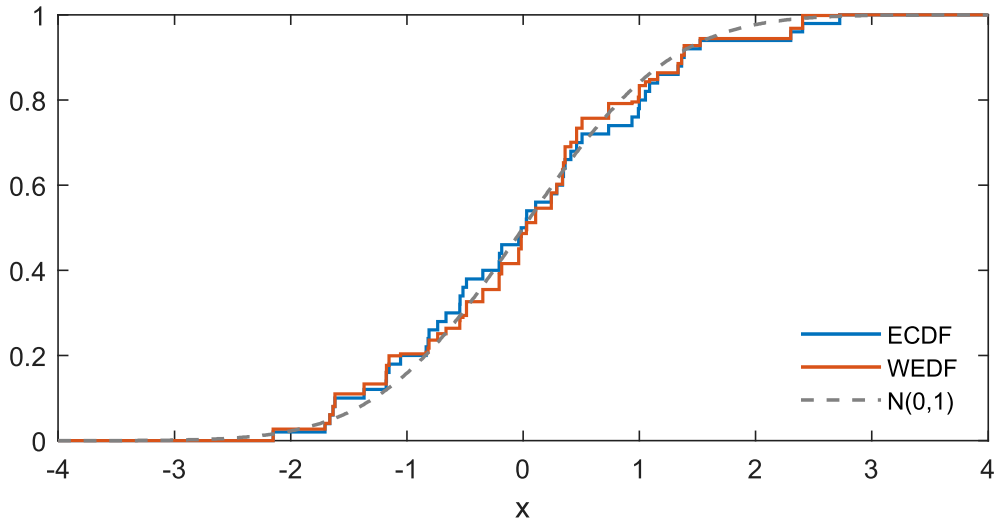
2.2 Modifikace testu pro vážená data

Popišme nyní na příkladu Kolmogorovova-Smirnovova testu modifikaci testovací statistiky pro účely porovnání rozdělení vážených dat, která byla navržena v [37, 38]. V dalším textu budeme na váhy nahlížet jako na náhodné veličiny, které mají konečnou střední hodnotu různou od nuly a konečný rozptyl. Protože klasické neparametrické testy homogenity pracují s empirickou distribuční funkcí (2.5), je potřeba ji nahradit jejím ekvivalentem pro vážená data. Pro tento účel tedy v souladu s [13] definujeme váženou empirickou distribuční funkci.

Definice 2.2.1. Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z F a ke každému prvku náhodného výběru je přiřazena kladná hodnota reprezentující jeho váhu $W_1, W_2, \ldots, W_n$. Dále označme $W = \sum_{j=1}^{n} W_j$ Pak definujeme *váženou empirickou distribuční funkci* (anglicky weighted empirical distribution function, zkráceně WEDF)

$$F_n^W(x) = \frac{1}{W} \sum_{j=1}^{n} W_j \, \mathbf{I}_{(-\infty,x]}(X_j), \quad \forall x \in \mathbb{R}. \tag{2.16}$$

Zdůrazněme, že pokud platí $W_j = 1$ pro $\forall j \in \{1, \ldots, n\}$ (tedy data jsou nevážená), tak WEDF (2.16) přechází v EDF (2.5). Porovnání EDF a WEDF pro 50 pozorování z normálního rozdělení N(0,1) s nezávislými váhami z rozdělení U$(0, 1)$ je znázorněno na obrázku 2.2.



Obrázek 2.2: Modrá křivka zobrazuje ECDF pro 50 pozorování z rozdělení N$(0, 1)$. Červená křivka je zobrazením WEDF pro tytéž pozorování z normálního rozdělení, ke kterým byly náhodně přiřazeny nezávislé váhy z uniformního rozdělení U$(0, 1)$.

Nyní pro test homogenity vážených dat uvažujme opět dva náhodné výběry, první $X_1, X_2, \ldots, X_n$ s váhami $W_1, W_2, \ldots, W_n$ označený $(\boldsymbol{X}, \boldsymbol{W})$ z rozdělení F^W a druhý $Y_1, Y_2, \ldots, Y_m$ s váhami $V_1, V_2, \ldots, V_m$ označený $(\boldsymbol{Y}, \boldsymbol{V})$ z rozdělení G^V. V případě KS testu vážených souborů nahradíme Kolmogorovu vzdálenost dvou ECDF $K_{n,m}$ Kolmogorovou vzdáleností dvou WEDF pro příslušné vážené soubory dat $(\boldsymbol{X}, \boldsymbol{W})$ a $(\boldsymbol{Y}, \boldsymbol{V})$

$$K_{n,m}^W = \sup_{x \in \mathbb{R}} \left| F_n^W(x) - G_m^V(x) \right|. \tag{2.17}$$

Druhou částí testovací statistiky KS testu, kterou potřebujeme modifikovat, je výraz $\sqrt{\frac{nm}{n+m}}$. Přímočarou volbou je nahrazení počtů pozorování n, m součty vah náhodných výběrů, při značení $W = \sum_{j=1}^{n} W_j$ a $V = \sum_{j=1}^{m} V_j$ to znamená, že

$$\sqrt{\frac{nm}{n+m}} \quad \text{nahradíme pomocí} \quad \sqrt{\frac{WV}{W+V}}. \tag{2.18}$$

Dle výsledků uvedených v [36] bylo toto zobecnění vyhodnoceno jako nestabilní, protože při numerickém ověření dodržení hladiny významnosti při užití modifikované testovací statistiky a rozdělení (2.12) závisel podíl zamítnutých testů na střední hodnotě a rozptylu vah. Druhým problémem je, že i při pouhém vynásobení vah konstantou se mění hodnota testovací statistiky a tím i spočtená p-hodnota, na základě které se rozhodujeme o zamítnutí či nezamítnutí H_0.

Vhodnější veličinou k nahrazení počtu pozorování u vážených dat je takzvaná *efektivní velikost vzorku*

$$n_e = \frac{\left(\sum_{j=1}^{n} W_j\right)^2}{\sum_{j=1}^{n} W_j^2} \approx n \frac{(\mathrm{E}\, W)^2}{\mathrm{E}\, W^2}, \tag{2.19}$$

která byla poprvé navržena v [44] a blíže popsána v [1]. Na rozdíl od prostého součtu vah n_e zohledňuje rozptyl vah, ale především je invariantní na škálování vah, takže spočtená p-hodnota nezávisí na vynásobení vah nenulovou konstantou.

Nahradíme-li klasickou ECDF (2.5) váženou distribuční funkcí WEDF (2.16) a počty pozorování efektivní velikostí vzorku, dostáváme vzorec modifikované testovací statistiky KS testu

$$\frac{n_e m_e}{n_e + m_e} \sup_{x \in \mathbb{R}} \left| F_n^W(x) - G_m^V(x) \right|. \tag{2.20}$$

Velmi důležité je podotknout, že pro tuto upravenou testovací statistiku není známo její asymptotické rozdělení, nemáme tedy ekvivalent ke vztahu (2.12), a nebyla dokázána ani stejnoměrná konvergence F_n^W jako u Glivenko-Cantelliho teorému (2.8) pro F_n.

Pokud uvažujeme váhy W nezávislé na pozorování X z rozdělení F, pak dle [37] pro F_n^W platí, že F_n^W je nestranným a silně konzistentním odhadem distribuční funkce F.

V [37, 38] bylo provedeno numerické ověření funkčnosti upraveného testu homogenity sledováním chyby I. druhu a následně také analýzou síly testu. Testována byla shodnost rozdělení pro dva soubory dat, jeden s váhami a druhý bez vah, jako ekvivalent k úloze testu shodnosti rozdělení MC simulací a reálných měření. Pro pozorování X bylo zvoleno rozdělení $\mathrm{N}(0, 1)$, nezávislé váhy prvního souboru dat byly generovány z rozdělení Beta nebo Gamma s různými parametry. Výsledky uvedené v [37] ukázaly, že pro zmíněné třídy vah lze modifikované testy označit za stabilní z pohledu chyby I. druhu, protože při předpokladu shodnosti rozdělení obou souborů byla chyba I. druhu přibližně stejná jako zvolená hladina významnosti $\alpha = 0,05$. Tyto výsledky naznačují, že asymptotické rozdělení modifikované statistiky (2.20) je nejspíš opravdu stejné nebo velmi blízké rozdělení standardní testovací statistiky (2.12).

2.3 Transformace vážených dat

Modifikace testovací statistiky pro vážená data, která byla popsána v předchozí části, je jednou z možností, jak lze přistoupit k problematice testování shodnosti rozdělení vážených souborů dat. V této části se zaměříme na odlišný přístup, vážený soubor dat nejprve transformujeme na nevážený soubor, aby bylo možné následně použít některý ze standardních testů homogenity, pro který je odvozeno asymptotické rozdělení testovací statistiky. V této práci jsou použity dvě metody pro transformaci vážených dat na nevážená, první z nich je metoda tzv. *přeuspořádání* (angl. re-arranging) představená v [13, 14], druhou pak generování neváženého souboru z jádrového odhadu pro vážená data.

2.3.1 Re-arranging (přeuspořádání)

Metoda přeuspořádání (angl. re-arranging) spočívá v nahrazení vážených pozorování specificky spočtenými váženými průměry, které vytvoří nový nevážený soubor dat. Tato transformace využívá informace obsažené ve váhách jednotlivých pozorování a zároveň je konstruovaná takovým způsobem, že počet nových pozorování N splňuje $N = \lfloor \sum_{j=1}^{n} w_j \rfloor$.

Předpokládejme konkrétní uspořádanou realizaci náhodného výběru $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ s váhami $w_1, w_2, \ldots, w_n$, které splňují, že pro každé $i \in \{1, \ldots, n\}$ platí $0 \leq w_i \leq 1$. Pro konstrukci prvního neváženého pozorování $y_{(1)}$ vezmeme k_1 nejmenší možný počet vážených pozorování $x_{(1)}, \ldots, x_{(k_1)}$ tak, aby

$$1 \leq \sum_{i=1}^{k_1} w_i < 2, \tag{2.21}$$

nutně tedy pro všechna $l < k_1$ platí

$$\sum_{i=1}^{l} w_i < 1. \tag{2.22}$$

Pro pozorování x_{k_1} zahrneme do výpočtu $y_{(1)}$ pouze tu část váhy w_{k_1}, která dorovná součet $\sum_{i=1}^{k_1-1} w_i$ na hodnotu 1. Zbylou reziduální část označíme

$$r_{k_1} = \sum_{i=1}^{k_1} w_i - 1. \tag{2.23}$$

První nevážené pozorování $y_{(1)}$ pak spočteme jako vážený průměr pozorování $x_{(1)}, \ldots, x_{(k_1-1)}$ a pozorování $x_{(k_1)}$ s váhou $w_{k_1} - r_{k_1}$, tedy

$$y_{(1)} = \frac{\sum_{i=1}^{k_1} x_{(i)} w_i - x_{(k_1)} r_{k_1}}{\sum_{i=1}^{k_1} w_i - r_{k-1}} = \sum_{i=1}^{k_1-1} x_{(i)} w_i + x_{(k_1)}(w_{k_1} - r_{k_1}), \tag{2.24}$$

protože výraz ve jmenovateli je díky definici rezidua (2.23) roven 1. Hodnotu x_{k_1} s váhou určenou reziduem r_{k_1} započítáme do váženého průměru při konstrukci dalšího neváženého pozorování $y_{(2)}$.

Obecné pozorování $y_{(j)}$ pro $j \in \{2, \ldots, N\}$ získáme tak, že nejprve najdeme nejmenší k_j takové, že

$$1 \leq \sum_{i=k_{j-1}+1}^{k_j} w_i + r_{k_{j-1}} < 2, \tag{2.25}$$

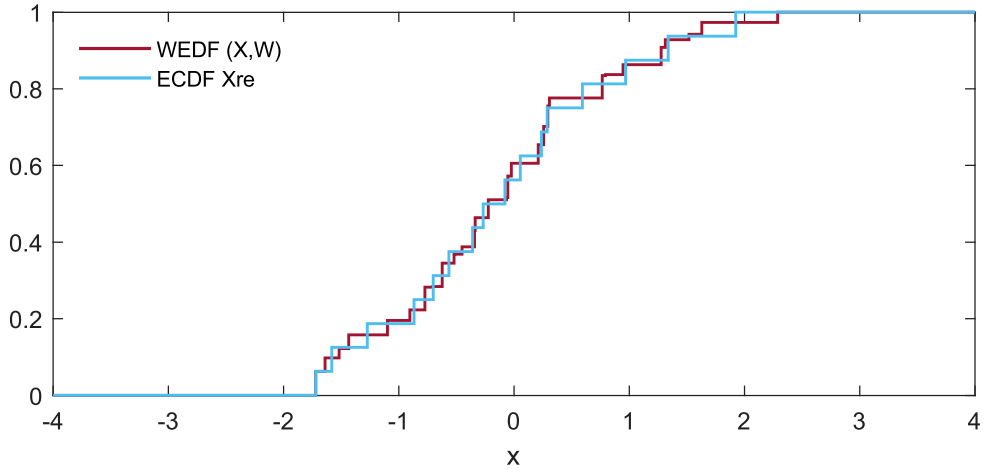
kde k součtu vah přičítáme reziduum $r_{k_{j-1}}$. Reziduum r_{k_j} pak definujeme jako

$$r_{k_j} = \sum_{i=k_{j-1}+1}^{k_j} w_i + r_{k_{j-1}} - 1 \tag{2.26}$$

a pozorování $y_{(j)}$ spočteme opět jako vážený součet

$$y_{(j)} = x_{(k_{j-1})} r_{k_{j-1}} + \sum_{i=k_{j-1}+1}^{k_j-1} x_{(i)} w_i + x_{(k_j)}(w_{k_j} - r_{k_j}). \tag{2.27}$$

Aby také pro poslední nevážené pozorování y_N vyšel součet vah $\sum_{i=k_{N-1}}^{n} w_i = 1$, je potřeba úplně na začátku přeškálovat všechny váhy tak, aby jejich součet byl roven $N = \lfloor \sum_{j=1}^{n} w_j \rfloor$. Graf 2.3 ilustruje rozdíl mezi WEDF pro vážený soubor dat a ECDF pro data získaná metodou re-arranging.



Obrázek 2.3: Červená křivka vykresluje WEDF pro 30 pozorování s váhami (X, W), kde $X \sim \mathrm{N}(0,1)$ a $W \sim \mathrm{U}(0,1)$. Modrá křivka zobrazuje ECDF pro nevážený dataset získaný metodou re-arranging.

2.3.2 Jádrové odhady

Jádrový odhad (kernel density estimate, zkráceně KDE) je neparametrická metoda odhadu hustoty pravděpodobnosti na základě souboru dat. Metoda KDE je jednoduše využitelná pro vážená data, proto je vhodným adeptem pro využití v kontextu problematiky testů homogenity vážených dat. V této podkapitole popíšeme klasický KDE a adaptivní KDE (zkráceně AKDE) pouze pro jednorozměrné odhady, definici pro vícerozměrný případ je možné nalézt v [34]. Tyto jádrové odhady později využijeme pro návrh nového přístupu ke zobecnění testů homogenity pro vážená data.

Definice 2.3.1. Uvažujme $X_1, X_2, \ldots, X_n$ *i.i.d.* náhodný výběr s hustotou pravděpodobnosti $f : \mathbb{R} \to \mathbb{R}_0^+$. Pak *jádrovým odhadem* (KDE) hustoty f v bodě $t \in \mathbb{R}$ nazveme

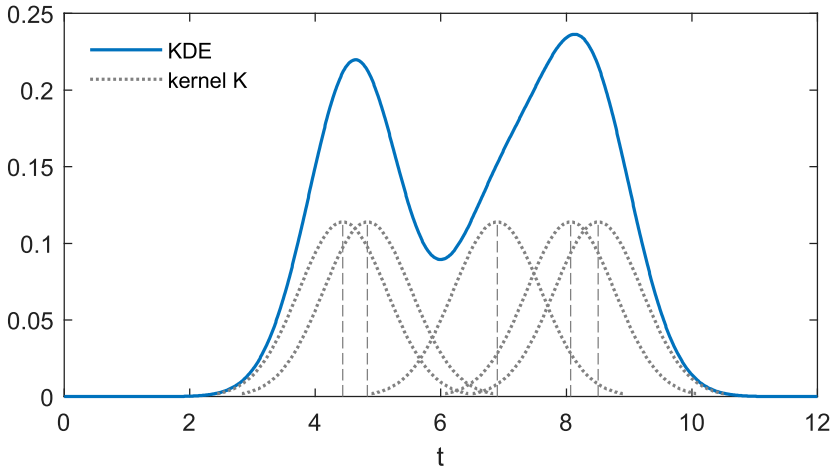
$$\hat{f}(t) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{t - X_j}{h}\right), \tag{2.28}$$

kde $h > 0$ nazveme *vyhlazovací parametr* (angl. bandwidth) a funkci $K : \mathbb{R} \to \mathbb{R}_0^+$ splňující $\int_{\mathbb{R}} K(t)\,\mathrm{d}t = 1$ nazveme *jádrem*.

Hodnota jádrového odhadu v bodě $t \in \mathbb{R}$ je tedy dána součtem hodnot jádra K posunutého do bodů $X_1, \ldots, X_n$ a přeškálovaného parametrem h, jak je znázorněno na obrázku 2.4. Jako jádro K se obvykle volí symetrická funkce a díky dalším požadavkům na funkci K uvedeným v definici 2.3.1 pro $\hat{f}$ platí

$$\int_{\mathbb{R}} \hat{f}(t)\,\mathrm{d}t = 1, \tag{2.29}$$

tedy výsledný odhad $\hat{f}$ je hustotou pravděpodobnosti. Tabulka 2.2 obsahuje obvyklé funkce volené jako jádro K.



Obrázek 2.4: Ilustrace principu konstrukce KDE pro 5 pozorování. Modrá křivka zobrazuje KDE, tečkované křivky odpovídají $\frac{1}{nh}K\left(\frac{t-X_j}{h}\right)$ pro $j \in \{1,\ldots,5\}$.

Tabulka 2.2: Příklady jader K a jejich eficiencí [34].

Název	$K(t)$		$eff\,(K)$
Epanechnikovo jádro	$\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right)$ 0	pro $\lvert t\rvert < \sqrt{5}$ jinak	1
Gaussovské jádro	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$		0.9512
Biweight	$\frac{15}{16}(1-t^2)^2$ 0	pro $\lvert t\rvert < 1$ jinak	0.9939
Triweight	$\frac{35}{32}(1-t^2)^3$ 0	pro $\lvert t\rvert < 1$ jinak	0.9867
Rectangular	$\frac{1}{2}$ 0	pro $\lvert t\rvert < 1$ jinak	0.9295
Triangular	$1-\lvert t\rvert$ 0	pro $\lvert t\rvert < 1$ jinak	0.9859

Chyba jádrového odhadu

Jádrový odhad závisí na naměřených pozorováních, ale také na funkci K a parametru šířky okna h, které jsou volené uživatelem. Určitá doporučení pro volbu jádra a parametru h lze odvodit analýzou střední kvadratické chyby jádrového odhadu. Předpokládejme tedy náhodný výběr $X_1, \ldots, X_n$ z rozdělení s hust. pravděp. f, která je spojitá. *Střední kvadratickou chybu* (mean square error) jádrového odhadu $\hat{f}$ v konkrétním bodě $t \in \mathbb{R}$ lze zapsat jako

$$\begin{aligned} \mathrm{MSE}_t(\hat{f}) &= \mathrm{E}\left(\hat{f}(t) - f(t)\right)^2 = \left(\mathrm{E}\,\hat{f}(t) - f(t)\right)^2 + \mathrm{E}\,(\hat{f}(t))^2 - \left(\mathrm{E}\,\hat{f}(t)\right)^2 \\ &= \left(\mathrm{Bias}\,\hat{f}(t)\right)^2 + \mathrm{Var}\,\hat{f}(t), \end{aligned} \tag{2.30}$$

kde Bias je označení pro vychýlení a Var značí rozptyl. Střední hodnoty jsou počítány přes náhodný výběr $X_1, \ldots, X_n$, z důvodu přehlednosti zápisu to není explicitně označeno.

Celkovou chybu jádrového odhadu ve srovnání s původní hustotou pravděpodobnosti f lze vyjádřit pomocí *integrované střední kvadratické chyby* (mean integrated square error)

$$\mathrm{MISE}(\hat{f}) = \mathrm{E} \int_{\mathbb{R}} \left(\hat{f}(t) - f(t)\right)^2 \mathrm{d}t, \tag{2.31}$$

kterou lze díky nezápornosti integrované funkce zapsat jako integrál ze střední kvadratické chyby (2.30)

$$\mathrm{MISE}(\hat{f}) = \int_{\mathbb{R}} \mathrm{MSE}_t(\hat{f})\,\mathrm{d}t = \int_{\mathbb{R}} \left(\mathrm{Bias}\,\hat{f}(t)\right)^2 \mathrm{d}t + \int_{\mathbb{R}} \mathrm{Var}\,\hat{f}(t)\,\mathrm{d}t. \tag{2.32}$$

Pro střední hodnotu jádrového odhadu, vyskytující se ve výrazu pro $\mathrm{Bias}\,\hat{f}(t)$, lze odvodit

$$\mathrm{E}\,\hat{f}(t) = \frac{1}{nh} \sum_{j=1}^{n} \mathrm{E}\left(K\left(\frac{t - X_j}{h}\right)\right) = \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{t-x}{h}\right) f(x)\,\mathrm{d}x. \tag{2.33}$$

Integrál v posledním výrazu obvykle není snadné spočíst. Z tohoto výrazu lze ale vidět, že střední hodnota jádrového odhadu v bodě t je dána konvolucí původní hustoty pravděpodobnosti f s jádrem K škálovaným pomocí parametru h. Hodnota $\hat{f}(t)$ je tedy dána hodnotou vyhlazené hustoty f a přičtením náhodné chyby $\varepsilon \in \mathbb{R}$.

Pro získání vhodnějšího vyjádření chyby MISE (2.32) vyjádříme přibližnou hodnotu vychýlení $\mathrm{Bias}\,\hat{f}(t)$ a rozptylu $\mathrm{Var}\,\hat{f}(t)$. Předpokládejme, že f má spojité derivace do potřebného řádu a že jádro K splňuje

$$\int_{\mathbb{R}} K(t)\,\mathrm{d}t = 1, \quad \int_{\mathbb{R}} tK(t)\,\mathrm{d}t = 0, \quad \int_{\mathbb{R}} t^2 K(t)\,\mathrm{d}t = k_2 \neq 0. \tag{2.34}$$

S využitím (2.33) lze pro $\mathrm{Bias}\,\hat{f}(t)$ psát

$$\begin{aligned} \mathrm{Bias}\,\hat{f}(t) &= \mathrm{E}\,\hat{f}(t) - f(t) = \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{t-x}{h}\right) f(x)\,\mathrm{d}x - f(t) \\ &= \int_{\mathbb{R}} K(y) f(t - hy)\,\mathrm{d}y - f(t) = \int_{\mathbb{R}} K(y)\left(f(t-hy) - f(t)\right)\mathrm{d}y. \end{aligned} \tag{2.35}$$

Dosazením Taylorova rozvoje $f(t-hy) = f(t) - hyf'(t) + \frac{1}{2}h^2y^2f''(t) + \cdots$ a s použitím (2.34) dostáváme

$$\mathrm{Bias}\,\hat{f}(t) = -hf'(t) \int_{\mathbb{R}} K(y)\,\mathrm{d}y + \frac{1}{2}h^2 f''(t) \int_{\mathbb{R}} y^2 K(y)\,\mathrm{d}y + \cdots \approx \frac{1}{2}h^2 k_2 f''(t) \tag{2.36}$$

za předpokladu, že h je dostatečně malé. Získáváme tak aproximaci

$$\int_{\mathbb{R}} \left(\operatorname{Bias}\hat{f}(t)\right)^2 \,\mathrm{d}t \approx \frac{1}{4}h^4 k_2^2 \int_{\mathbb{R}} f''(x)^2 \,\mathrm{d}x. \tag{2.37}$$

Obdobným způsobem lze získat přibližné vyjádření rozptylu $\operatorname{Var}\hat{f}(t)$. S využitím předpokladu nezávislosti a shodnosti rozdělení $X_1, \ldots, X_n$ pro rozptyl $\hat{f}(t)$ platí

$$\begin{aligned}
\operatorname{Var}\hat{f}(t) &= \operatorname{Var}\left(\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{t-X_j}{h}\right)\right) = \frac{1}{nh^2}\operatorname{Var} K\left(\frac{t-X_1}{h}\right) \\
&= \frac{1}{nh^2}\left[\mathrm{E}\, K\left(\frac{t-X_1}{h}\right)^2 - \left(\mathrm{E}\, K\left(\frac{t-X_1}{h}\right)\right)^2\right] \\
&= \frac{1}{nh^2}\left[\int_{\mathbb{R}} K\left(\frac{t-x}{h}\right)^2 f(x)\,\mathrm{d}x - \left(\int_{\mathbb{R}} K\left(\frac{t-x}{h}\right) f(x)\,\mathrm{d}x\right)^2\right].
\end{aligned} \tag{2.38}$$

Druhý z integrálů se vyskytuje také v rovnici (2.35), lze tedy psát

$$\begin{aligned}
\operatorname{Var}\hat{f}(t) &= \frac{1}{nh^2}\int_{\mathbb{R}} K\left(\frac{t-x}{h}\right)^2 f(x)\,\mathrm{d}x - \frac{1}{n}\left(\operatorname{Bias}\hat{f}(t) + f(t)\right)^2 \\
&= \frac{1}{nh}\int_{\mathbb{R}} K(y)^2 f(t-hy)\,\mathrm{d}y - \frac{1}{n}\left(f(t) + O(h^2)\right)^2.
\end{aligned} \tag{2.39}$$

Použitím Taylorova rozvoje $f(t-hy)$ stejně jako v (2.36) získáváme

$$\operatorname{Var}\hat{f}(t) = \frac{1}{nh}\int_{\mathbb{R}} K(y)^2\left[f(t) - hyf'(t) + \frac{1}{2}h^2y^2f''(t) - \cdots\right]\mathrm{d}y - \frac{1}{n}\left(f(t) + O(h^2)\right)^2. \tag{2.40}$$

Za předpokladu velkého počtu pozorování n a malé šířky okna h získáváme aproximaci

$$\operatorname{Var}\hat{f}(t) \approx \frac{1}{nh} f(t)\int_{\mathbb{R}} K(y)^2\,\mathrm{d}y. \tag{2.41}$$

Integrál z rozptylu vyskytující se v (2.32) lze tedy aproximovat výrazem

$$\int_{\mathbb{R}} \operatorname{Var}\hat{f}(t)\,\mathrm{d}t \approx \frac{1}{nh}\int_{\mathbb{R}} K(y)^2\,\mathrm{d}y. \tag{2.42}$$

Dosazením (2.37) a (2.42) do (2.32) získáváme aproximaci MISE

$$\operatorname{MISE}(\hat{f}) \approx \frac{1}{4}h^2k_2^2\int_{\mathbb{R}} f''(x)\,\mathrm{d}x + \frac{1}{nh}\int_{\mathbb{R}} K(x)^2\,\mathrm{d}x. \tag{2.43}$$

Z výrazu vidíme, že při zvětšení šířky okna h se zvětší také vychýlení neboli systematická chyba KDE. Naopak zmenšením h dojde ke zvětšení rozptylu jádrového odhadu, tedy ke zvětšení náhodné chyby. Obě složky chyby MISE tedy nelze minimalizovat současně a volba vhodného parametru h je o vybalancování systematické a náhodné chyby KDE.

Volba parametru h a jádra K

Z aproximace (2.43) lze odvodit vzorec pro parametr h optimální z hlediska MISE($\hat{f}$)

$$h_{opt} = n^{-\frac{1}{5}} k_2^{\frac{2}{5}} \left(\int_{\mathbb{R}} K(t)^2 \, \mathrm{d}t \right)^{\frac{1}{5}} \left(\int_{\mathbb{R}} f''(x)^2 \, \mathrm{d}x \right)^{-\frac{1}{5}}. \tag{2.44}$$

S rostoucím počtem pozorování n tedy klesá h_{opt}. Stejně tak pro hustotu f, která více fluktuuje, je vhodnější menší hodnota parametru h, což vyjadřuje člen obsahující $f''(x)^2$.

Dosazením h_{opt} optimální šířky okna pro dané jádro K a hustotu f do vyjádření MISE($\hat{f}$) získáváme

$$\mathrm{MISE}(\hat{f}) = \frac{5}{4} n^{-\frac{4}{5}} C(K) \left(\int_{\mathbb{R}} f''(x)^2 \, \mathrm{d}x \right)^{\frac{1}{5}}, \quad C(K) = k_2^{\frac{2}{5}} \left(\int_{\mathbb{R}} K(t)^2 \, \mathrm{d}t \right)^{\frac{4}{5}}. \tag{2.45}$$

Chybu MISE($\hat{f}$) lze tedy minimalizovat také vhodnou volbou jádra K. Optimálním jádrem z tohoto pohledu je tzv. *Epanechnikovo jádro*

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5} t^2\right) & \text{pro } |t| \leq 5, \\ 0 & \text{jinak.} \end{cases} \tag{2.46}$$

Pro porovnání různých funkcí jádra K s Epanechnikovým jádrem K_e se pak zavádí tzv. *eficience*

$$eff(K) = \left(\frac{C(K_e)}{C(K)} \right)^{\frac{5}{4}}. \tag{2.47}$$

Hodnoty eficiencí pro různá jádra jsou uvedeny v tabulce 2.2. Pro všechna uvedená jádra jsou hodnoty eficiencí blízké jedné, volba jádra tedy hraje spíše menší roli z hlediska integrované střední kvadratické chyby.

Větší vliv na chybu jádrového odhadu MISE($\hat{f}$) má tedy volba parametru h, který ale závisí na neznámé hustotě f. Jednou z možností, jak zvolit hodnotu parametru, je jeho výpočet pro referenční hustoty f. Konkrétně pro data z normálního rozdělení $\mathrm{N}(\mu, \sigma^2)$ a Gaussovské jádro platí

$$h_{opt} = 1,06 \cdot \sigma n^{-\frac{1}{5}}. \tag{2.48}$$

Pro unimodální rozdělení s těžšími chvosty nebo bimodální rozdělení je vhodnější volbou

$$h_{opt} = 0,9 \cdot A n^{-\frac{1}{5}} \text{ pro } A = \min\{\sigma, IQR/1,34\}, \tag{2.49}$$

kde IQR označuje interkvartilové rozpětí $x_{0,75} - x_{0,25}$. V obou případech v praxi při výpočtu h_{opt} dosazujeme odhady rozptylu, respektive IQR.

Adaptivní jádrový odhad

Klasický jádrový odhad z definice 2.3.1 používá stejnou hodnotu parametru h pro celý rozsah hodnot. Tato vlastnost je problematická především pro data z rozdělení s těžšími chvosty. Máme-li náhodný výběr $X_1, \ldots, X_n$ s hustotou pravděp. f, pak v oblastech s větší frekvencí pozorování by bylo vhodné použít menší hodnotu h, aby nedošlo k přílišnému vyhlazení odhadu a ztrátě důležité informace o hustotě pravděp. f. Menší hodnota h

pak ale může způsobovat problémy v oblastech s řídkým výskytem pozorování, kde bude odhad nepřiměřeně fluktuovat.

Adaptivní jádrový odhad (angl. adaptive kernel density estimate, AKDE) se snaží tento problém řešit uzpůsobením šířky okna h podle četnosti výskytu pozorování daného datasetu v okolí konkrétního bodu. Prvním krokem je konstrukce pilotního KDE $\hat{f}_{pilot}$ s parametrem h_{pilot}, po kterém požadujeme pouze $\hat{f}_{pilot}(X_j) > 0$ pro $\forall j \in \{1, \ldots, n\}$. S použitím $\hat{f}_{pilot}$ spočteme *lokální vyhlazovací faktor*

$$\lambda_j = \left(\frac{\hat{f}_{pilot}(X_j)}{g} \right)^{-\alpha} \quad \forall j \in \{1, \ldots, n\}, \text{ kde } g = \sqrt[n]{\prod_{j=1}^{n} \hat{f}_{pilot}(X_j)}. \tag{2.50}$$

Výraz g je geometrickým průměrem hodnot pilotního odhadu pro měřená pozorování. Parametr senzitivity $\alpha \in (0, 1]$ upravuje citlivost šířky okna na četnost lokálního výskytu pozorování, protože s pomocí λ_j definujeme pro každé pozorování X_j vlastní parametr šířky okna $h_j = \lambda_j h_{pilot}$. Nakonec konstruujeme AKDE

$$\hat{f}(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_j} K\left(\frac{t - X_j}{h_j} \right), \quad \forall t \in \mathbb{R}. \tag{2.51}$$

Dle [2] není metoda AKDE příliš citlivá na kvalitu pilotního odhadu $\hat{f}_{pilot}$, proto je možné jako h_{pilot} zvolit například hodnotu optimální pro normální rozdělení (2.48). Pro parametr citlivosti je doporučenou volbou $\alpha = \frac{1}{2}$.

2.3.2.1 Jádrový odhad pro vážená data

Zobecnění jádrového odhadu KDE (2.28) i AKDE (2.51) pro vážená data je přímočaré. Uvažujme náhodný výběr $X_1, \ldots, X_n$ z rozdělení s hustotou pravděp. f a příslušné váhy $W_1, \ldots, W_n$. *Vážený jádrový odhad* (angl. weighted KDE, zkráceně WKDE) definujeme jako

$$\hat{f}(t) = \frac{1}{h \sum_{j=1}^{n} W_j} \sum_{j=1}^{n} W_j \, K\left(\frac{t - X_j}{h} \right), \quad \forall t \in \mathbb{R}. \tag{2.52}$$

Analogickým způsobem lze provést zobecnění adaptivního jádrového odhadu na *vážený AKDE* (angl. weighted AKDE, zkráceně WAKDE)

$$\hat{f}(t) = \frac{1}{\sum_{j=1}^{n} W_j} \sum_{j=1}^{n} \frac{W_j}{h_j} \, K\left(\frac{t - X_j}{h_j} \right), \quad \forall t \in \mathbb{R}. \tag{2.53}$$

Pokud jsou všechny váhy rovny jedné, přechází WKDE na klasický KDE a WAKDE na AKDE.

Generování z KDE

Jádrový odhad chceme využít v testech homogenity k transformaci váženého souboru dat na nevážená pozorování. Nejprve tedy použijeme vážená pozorování ke konstrukci WKDE (2.52), resp. WAKDE (2.53). Následně ze získaného odhadu $\hat{f}$ budeme generovat nový soubor nevážených pozorování $X'_1, \ldots, X'_p$.

První možností, jak lze generovat nová pozorování z jádrového odhadu, je tzv. **generování z náhodného výběru** [22]. Pro KDE s jádrem K a šířkou okna h sestává postup ze třech kroků.

1. Vybereme náhodně jednu hodnotu z náhodného výběru $X_1, \ldots, X_n$, tedy generujeme náhodný index $I \sim \mathrm{U}(\{1, \ldots, n\})$.

2. Generujeme náhodné číslo ε z rozdělení s hustotou pravděpodobnosti K, neboli $\varepsilon \sim K$.

3. Definujeme $X' = X_I + h\varepsilon$.

Pro generování pozorování z AKDE je postup analogický, pouze ve třetím kroku spočteme nové nevážené pozorování s použitím upravené šířky okna, tedy $X' = X_I + h_I\varepsilon$. Pokud pracujeme s váženými daty a WKDE, resp. WAKDE, náhodný index I v prvním kroku negenerujeme z diskrétního rovnoměrného rozdělení, ale ze zobecněného Bernoulliho rozdělení, kde pravděpodobnost vybrání indexu $j \in \{1, \ldots, n\}$ odpovídá váze původně přiřazené pozorování X_j.

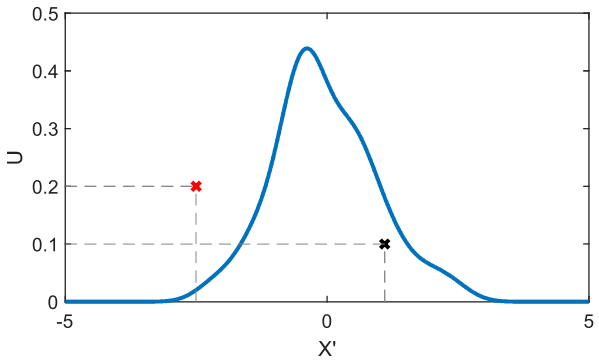
Druhou možností je jednoduché **generování z křivky KDE**. Pro odhad s jádrem K, které má neomezený support, je potřeba určit interval, ze kterého budeme nová nevážená pozorování generovat. Pokud víme, že skutečné rozdělení f má $\operatorname{supp} f = \mathbb{R}$, můžeme pro generování nových hodnot určit například interval $[-5\hat{\sigma}, 5\hat{\sigma}]$, kde $\hat{\sigma}$ je odhad směrodatné odchylky. Generování z křivky lze pak rozdělit do třech kroků, grafické znázornění je k dispozici na obrázku 2.5.

1. Generujeme $X' \sim \mathrm{U}([-5\hat{\sigma}, 5\hat{\sigma}])$, respektive $X' \sim \mathrm{U}(\operatorname{supp} f)$ pokud je $\operatorname{supp} f$ omezený.

2. Generujeme $U \sim \mathrm{U}([0, \max_{t\in\mathbb{R}} \hat{f}(t)])$.

3. Pokud $U > \hat{f}(X')$, hodnoty U a X' zahazujeme a opakujeme kroky 1. a 2. Pokud platí $U \leq \hat{f}(X')$, přidáme hodnotu X' do nově vznikajícího souboru nevážených pozorování.

Všechny tři kroky opakujeme, dokud nemáme požadovaný počet nevážených pozorování. Tento postup je sice pro generování z jádrového odhadu přesnější, nevýhodou je jeho výrazně vyšší časová náročnost a případná nutnost omezení intervalu, ze kterého hodnoty generujeme, pokud má zvolené jádro neomezený support.



Obrázek 2.5: Ilustrace principu generování nových pozorování z křivky KDE. Pouze černě označený bod bude přidán do nového souboru pozorování.

Kapitola 3

Numerické simulace testů homogenity

Způsoby testování homogenity vážených dat navržené v předchozí kapitole byly podrobeny analýze provedením numerických simulací. Modifikace testovací statistiky pro vážená data popsaná v sekci 2.2 byla zkoumána už v diplomové práci [37] pro případ testování souboru dat s váhami proti souboru dat bez vah, ve kterých hodnoty pozorování $\boldsymbol{X}$ byly generovány z normálního rozdělení $\mathrm{N}(0,1)$ a váhy $\boldsymbol{W}$ z rozdělení Gamma nebo Beta. Metoda re-arrangingu popsaná v sekci 2.3.1 byla rovněž zkoumána pouze pro případ porovnávání váženého souboru dat s neváženým pro váhy z rozdělení $\mathrm{Beta}(2,5)$ [14].

Záměrem této kapitoly je blíže prozkoumat rozdíly mezi různými přístupy k testování homogenity vážených dat a jejich limity, a to pro více rodin rozdělení hodnot pozorování a vah. Porovnávány byly testy se statistikou modifikovanou pro vážená data, transformace vážených dat metodou re-arranging a nově také možnost využití vážených jádrových odhadů pro generování nevážených pozorování. Tyto přístupy byly navíc testovány na obecnější úloze zkoumání shodnosti rozdělení dvou vážených souborů dat. Pro test s modifikovanou statistikou a test využívající re-arranging bylo numerické ověření jejich funkčnosti rozšířeno na pět vybraných rodin rozdělení pozorování $\boldsymbol{X}$ ve srovnání s předchozími publikacemi.

3.1 Popis simulací

V numerických simulacích testů homogenity byly porovnávány dva vážené soubory pozorování $(\boldsymbol{x}, \boldsymbol{w})$ a $(\boldsymbol{y}, \boldsymbol{v})$, tedy dvě realizace náhodných výběrů. Pozorování $\boldsymbol{x}$, respektive $\boldsymbol{y}$, byla generována z pěti různých rozdělení zahrnujících rozdělení s nosičem $\mathbb{R}$ i $\mathbb{R}^+$ a s různě těžkými chvosty. Příslušné hustoty pravděpodobnosti a konkrétní hodnoty parametrů jsou k dispozici na obrázku 3.1. Váhy byly generovány nezávisle na pozorováních z rozdělení Beta s různými parametry a z uniformního rozdělení $\mathrm{U}(0,1)$, hustoty pravděpodobnosti použitých rozdělení a konkrétní hodnoty parametrů jsou zobrazeny v grafech 3.2.

Numerická simulace byla provedena opakováním následujících kroků.

1. Generování dvou souborů pozorování $\boldsymbol{x} = (x_1, \ldots, x_s)$ a $\boldsymbol{y} = (y_1, \ldots, y_s)$ ze stejného rozdělení, kde oba soubory obsahují stejný počet pozorování s.

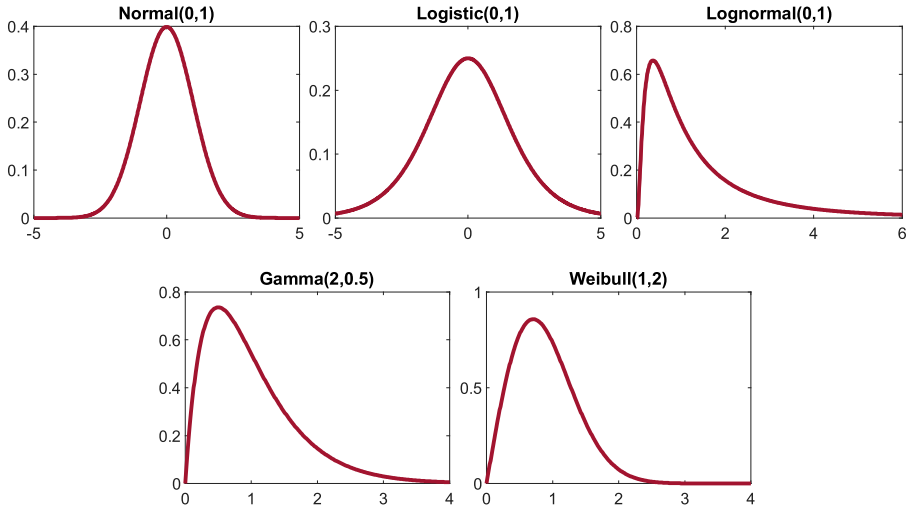
2. Generování dvou souborů vah $\boldsymbol{w} = (w_1, \ldots, w_s)$ a $\boldsymbol{v} = (v_1, \ldots, v_s)$ opět ze shodného rozdělení, váhy jsou nezávislé na pozorováních $\boldsymbol{x}$, $\boldsymbol{y}$.

3. Provedení vybrané varianty testu.

4. Zaznamenání rozhodnutí o zamítnutí nebo nezamítnutí hypotézy H_0, že oba vážené soubory dat $(\boldsymbol{x}, \boldsymbol{w})$ a $(\boldsymbol{y}, \boldsymbol{v})$ pochází ze stejného rozdělení při pevně zvolené hladině významnosti $\alpha = 0,05$. Dále byla zaznamenána také spočtená p-hodnota, na základě které bylo rozhodnutí testu učiněno.

Všechny výše popsané kroky byly provedeny v k opakováních. Vzhledem ke shodnému způsobu generování vážených souborů $(\boldsymbol{x}, \boldsymbol{w})$ a $(\boldsymbol{y}, \boldsymbol{v})$ víme, že hypotéza H_0: *oba soubory pochází ze stejného rozdělení* přibližně platí (v rámci kvality generátoru pseudonáhodných čísel). Pokud máme správně definovaný test naladěný na hladinu významnosti α a hypotéza H_0 platí, pak vzhledem k definici kritického oboru testu by mělo dojít k chybnému zamítnutí H_0 s pravděpodobností α.

Pokud generování dvou souborů dat ze stejného rozdělení a testování jejich homogenity opakujeme k-krát, pak očekáváme, že $\frac{\text{počet zamítnutí } H_0}{k} \approx \alpha$. Při zkoumání funkčnosti různých přístupů k testování vážených dat se proto v prvé řadě zaměříme na podíl zamítnutých testů ku provedeným testům, který odpovídá odhadu pravděpodobnosti chyby I. druhu. Simulace byly prováděny pro $k = 10\,000$ opakování, oba testované soubory měly vždy stejný počet pozorování s , konkrétně byly porovnávány datasety o velikostech $s \in \{500, 1000, 1500, 2000, 2500, 3000, 3500\}$.

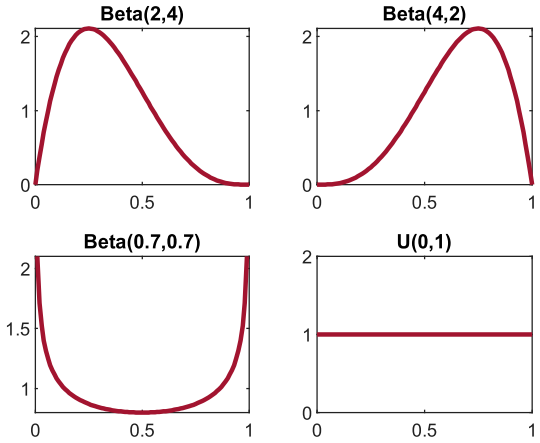


Obrázek 3.1: Hustoty pravděpodobnosti pro rozdělení náhodné veličiny X, resp. Y.

3.1.1 Výsledky simulací

Numerickým simulacím byly nejprve podrobeny čtyři varianty testů:

1. KS test se statistikou modifikovanou pro vážená pozorování, ve které je počet pozorování nahrazen efektivní velikostí vzorku (2.20),



Obrázek 3.2: Hustoty pravděpodobnosti pro rozdělení vah W, resp. V.

2. transformace váženého souboru dat na nevážený metodou re-arranging a následné testování homogenity pro nevážená data klasickým KS testem,

3. transformace váženého souboru dat na nevážený použitím WKDE (2.52) a generováním z náhodného výběru,

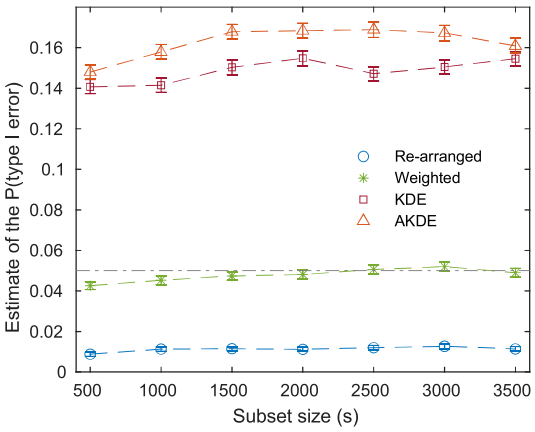
4. transformace váženého souboru dat na nevážený použitím WAKDE (2.53) a generováním z náhodného výběru.

Pro WKDE a WAKDE bylo zvoleno Gaussovské jádro K, které je vhodné vzhledem k použití normálního a logistického rozdělení pro testovaná pozorování a potřebě generovat z K náhodné hodnoty ε při vytváření neváženého datasetu. Parametr šířky okna h byl volen dle vzorce (2.48), respektive (2.49). Parametr citlivosti pro WAKDE byl nastaven na doporučenou hodnotu $\alpha_{\mathrm{AKDE}} = \frac{1}{2}$.

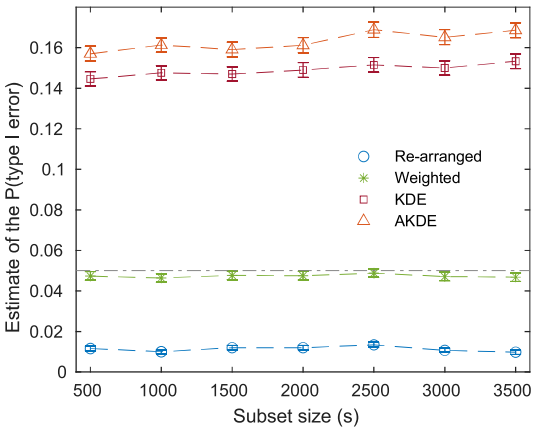
První výsledky odhadu pravděpodobnosti chyby I. druhu pro testování homogenity dvou vážených souborů jsou zobrazeny v grafech 3.3. Jednotlivé grafy se liší rozdělením použitým pro hodnoty pozorování $\boldsymbol{x}$, resp. $\boldsymbol{y}$, váhy byly ve všech případech generovány z rozdělení Beta$(4, 2)$. Výsledky pro různá rozdělení jsou velmi podobné, ve všech případech pozorujeme, že testy se statistikou modifikovanou pro vážená data (značené jako *weighted*) mají odhad pravděpodobnosti chyby I. druhu přibližně na hladině významnosti $\alpha = 0,05$. Při předchozím použití modifikovaných testů pro případ váženého souboru s neváženým bylo jejich fungování ověřeno numerickou simulací pouze pro soubory dat o 1000 pozorováních [37]. Nyní z grafů 3.3 vidíme, že testy fungují stejně dobře i pro dva vážené soubory o jiných počtech pozorování a pro odlišné typy distribucí.

Odhadnutá pravděpodobnost chyby I. druhu pro metodu re-arranging je pro všechna rozdělení a použité počty pozorování rovna přibližně 0,01, což je výrazně pod zvolenou hladinou významnosti. Tento výsledek znamená, že při použití metody re-arranging se test dopouští chybného zamítnutí hypotézy H_0 méně často, v takovém případě se ale dá očekávat horší síla testu, tedy schopnost správně zamítnout H_0, když neplatí.

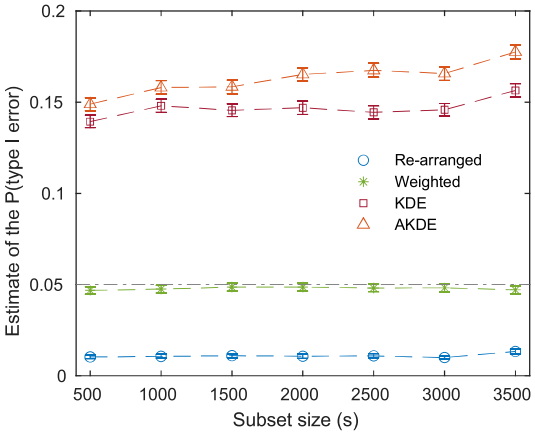
Při použití WKDE nebo WAKDE pro odhad hustoty pravděpodobnosti na váženém souboru dat a následném generování nevážených pozorování z tohoto odhadu je odhadnutá



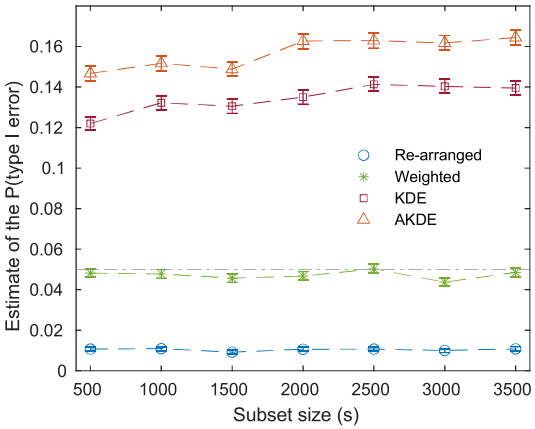
(a) Normální rozdělení



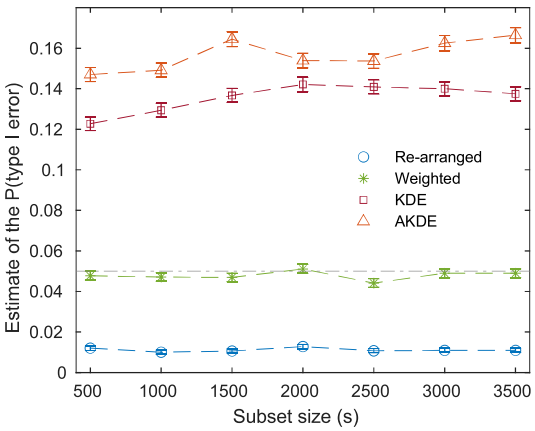
(b) Logistické rozdělení



(c) Lognormalní rozdělení



(d) Gamma rozdělení



(e) Weibullovo rozdělení

Obrázek 3.3: Odhad pravděpodobnosti chyby I. druhu pro testy dvou vážených souborů o s pozorováních ze stejného rozdělení. Pozorování $\boldsymbol{x}$, resp. $\boldsymbol{y}$ byla generována z rozdělení uvedeného v popisu grafu. Váhy byly generovány z rozdělení Beta$(4, 2)$. (*Re-arranged:* test s transformací dat pomocí metody re-arranging, *Weighted:* test s modifikovanou KS statistikou, *KDE:* test s generováním nevážených pozorování z WKDE, *AKDE*: test s generováním neváž. pozorování z WAKDE.)

pravděpodobnost chyby I. druhu přibližně trojnásobná ve srovnání s hladinou významnosti α, v grafech 3.3 příslušné křivky značené jako KDE a AKDE. Výsledky jsou velmi podobné pro všechna testovaná rozdělení, přestože jádro K a parametr h byly voleny optimálně pro data z normálního rozdělení z hlediska chyby MISE (2.32). Výsledky simulací pro váhy generované z rozdělení Beta s odlišnými parametry a váhy z uniformního rozdělení jsou k dispozici v příloze A.1.

Sledování odhadu pravděpodobnosti chyby I. druhu je založené na rozhodnutí o zamítnutí nebo nezamítnutí hypotézy H_0 při jedné pevně zvolené hladině významnosti α. Bližší pohled na chování testů homogenity poskytuje distribuční funkce p-hodnot. Považujme na chvíli p-hodnotu za náhodnou veličinu označenou V a definovanou pro pravostrannou variantu testu vztahem $V = 1 - F_T(T)$, kde F_T je distribuční funkce testovací statistiky T za platnosti hypotézy H_0. Pak pro distribuční funkci V platí

$$F_V(p) = \mathrm{P}(V \leq p) = \mathrm{P}(1 - F_T(T) \leq p) = \mathrm{P}(F_T(T) \geq 1 - p). \tag{3.1}$$

Za předpokladu spojitosti F_T a existence kvantilové funkce F_T^{-1} lze psát

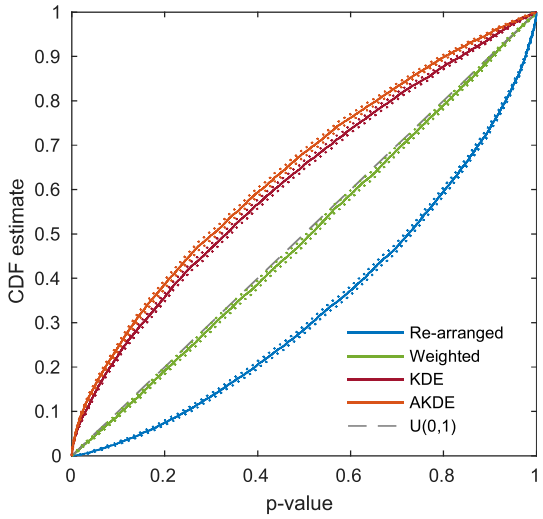
$$\begin{aligned} F_V(p) &= \mathrm{P}(T \geq F_T^{-1}(1-p)) = 1 - \mathrm{P}(T \leq F_T^{-1}(1-p)) \\ &= 1 - F_T(F_T^{-1}(1-p)) = p. \end{aligned} \tag{3.2}$$

Za platnosti H_0 se tedy náhodná veličina V reprezentující p-hodnotu testu řídí uniformním rozdělením $\mathrm{U}(0,1)$. Pro odvození výše ve skutečnosti stačí, aby F_T byla spojitá, pokud inverzní funkci F_T^{-1} nahradíme kvantilovou funkcí $F_T^{\leftarrow}(y) = \inf\{t \mid F_T(t) \geq y\}$.

Na obrázku 3.4 je vynesená empirická distribuční funkce získaná při simulacích testů homogenity pro dva soubory s hodnotami pozorování z rozdělení $\mathrm{N}(0,1)$ a váhami z rozdělení $\mathrm{Beta}(4,2)$ při počtu dat $s = 1000$. ECDF p-hodnot pro jednotlivé varianty testů koresponduje s grafy odhadů pravděpodobnosti chyby I. druhu. Pro test s modifikovanou KS statistikou vykreslená ECDF p-hodnot dobře kopíruje diagonální linii, která odpovídá uniformnímu rozdělení. Pro test používající re-arranging leží ECDF p-hodnot pod diagonálou, rozdělení je tedy vychýlené směrem k vyšším hodnotám, což odpovídá nízké hodnotě odhadu pravděpodobnosti chyby I. druhu. Naopak pro testy používající jádrové odhady je křivka ECDF nad diagonálou, rozdělení p-hodnot je tedy naopak vychýleno k nižším hodnotám a k zamítaní H_0 tak dochází častěji. Obdobné grafy pro další rodiny rozdělení pozorování jsou k dispozici v příloze A.2, závěry jsou pro všechna zkoumaná rozdělení analogické.

Jak již bylo zmíněno, nízká hodnota pravděpodobnosti chyby I. druhu u testu s metodou re-arranging nepředstavuje za platnosti hypotézy H_0 problém, ale dá se u něj očekávat nízká síla testu, tedy že za platnosti H_1 bude častěji docházet k chybnému nezamítnutí hypotézy H_0 o shodnosti rozdělení. Z tohoto důvodu byly provedeny také simulace pro získání odhadu síly testu.

Opět byly proti sobě testovány dva vážené soubory dat $(\boldsymbol{x}, \boldsymbol{w})$ a $(\boldsymbol{y}, \boldsymbol{v})$, ale část pozorování druhého datasetu $(\boldsymbol{y}, \boldsymbol{v})$ byla generována z jiného rozdělení než data $(\boldsymbol{x}, \boldsymbol{w})$. V prvním souboru dat byla všechna pozorování $\boldsymbol{x}$ generována z rozdělení $\mathrm{N}(0,1)$ a všechny váhy $\boldsymbol{w}$ z rozdělení $\mathrm{Beta}(4,2)$. Pro druhý soubor pozorování byla nejprve zvolena *míra znečištění* γ. Následně pro množinu $(\boldsymbol{y}, \boldsymbol{v})$ bylo $(100 - \gamma)\,\%$ pozorování a vah generováno ze stejných rozdělení jako pro soubor $(\boldsymbol{x}, \boldsymbol{w})$ a $\gamma\,\%$ pozorování tvořilo tzv. znečištění. Pro znečišťující data byly váhy nastaveny na pevnou hodnotu $\mathrm{E}\, W_1 = 0,67$ odpovídající střední hodnotě



Obrázek 3.4: Vykreslení empirické distribuční funkce p-hodnot včetně 95% konfidenčních intervalů (tečkované čáry) pro pozorování z rozdělení $\mathrm{N}(0, 1)$ a váhy z $\mathrm{Beta}(4, 2)$. Testované vážené soubory měly rozsah $s = 1000$.

rozdělení $\mathrm{Beta}(4, 2)$ a hodnoty pozorování generovány z normálního rozdělení s různými parametry, konkrétně $\mathrm{N}(0.5, 1)$, $\mathrm{N}(1, 1)$ a $\mathrm{N}(0.5, 1.5)$. Celkový počet pozorování v každém souboru byl shodně $s = 1000$.

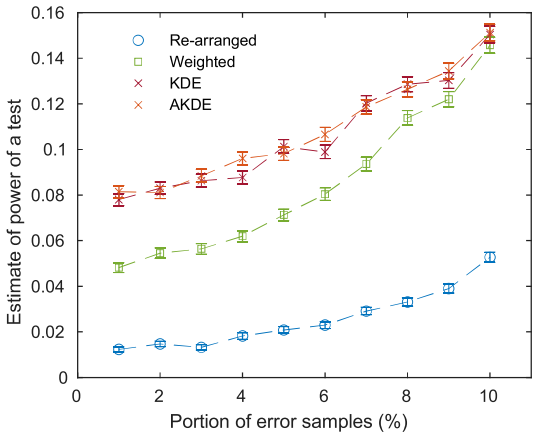
Síla testu byla následně odhadována jako $\frac{\text{počet zamítnutí } H_0}{k}$, kde počet simulací byl opět $k = 10\,000$. Výsledky těchto simulací jsou zobrazeny v grafech 3.5 pro hodnoty míry znečištění $\gamma \in \{1\,\%, 2\,\%, \ldots, 10\,\%\}$. V souladu s očekáváními byla nejnižší hodnota odhadu síly testu získána pro test používající re-arranging.

Zajímavé je srovnání testu s modifikovanou statistikou a testů využívajících jádrové odhady. Při nízké míře znečištění nejprve pozorujeme vyšší sílu testu pro varianty s jádrovými odhady. Vzhledem k tomu, že nízká pravděpodobnost chyby I. druhu a vysoká síla testu jsou v principu protichůdné jevy, tak vyšší síla testů s jádrovými odhady je zapříčiněná vyšší pravděpodobností chyby I. druhu (častější zamítání H_0, přestože platí). S rostoucí mírou znečištění druhého souboru dat roste odhad síly testu s modifikovanou statistikou rychleji než pro testy s jádrovými odhady. Při 10 % znečišťujícího rozdělení byla síla testu s modifikovanou statistikou minimálně na stejné úrovni, jako pro testy s jádrovými odhady. Ze zkoumaných variant tedy nejlépe fungují testy homogenity s modifikovanou statistikou a to jak z hlediska pravděpodobnosti chyby I. druhu, tak z hlediska síly testu.

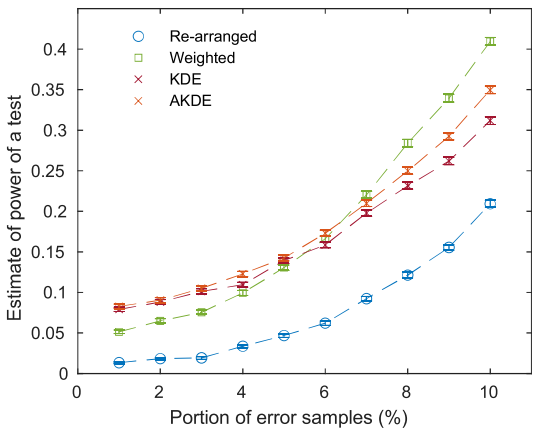
Vzhledem k vyššímu odhadu $\mathrm{P}(\mathit{chybaI.druhu})$ u testů s jádrovými odhady se dá očekávat, že po doladění těchto testů na správnou hladinu $\alpha = 0,05$ v síle testu překonají modifikovaný vážený test.

3.1.2 Použití jádrových odhadů

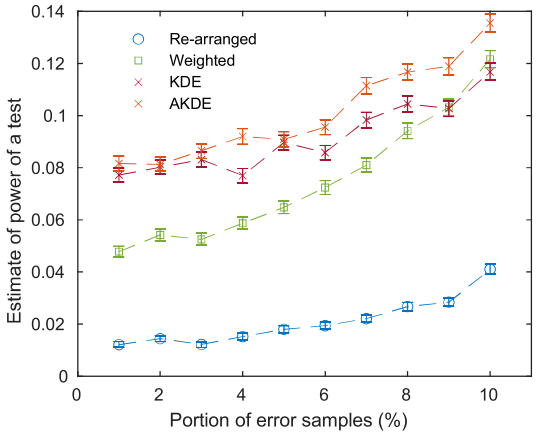
Vysoká hodnota odhadu pravděpodobnosti chyby I. druhu u testů využívajících jádrové odhady může být zapříčiněna buď způsobem generování pozorování z jádrového odhadu



(a) Znečišťující rozdělení $\mathrm{N}(0.5, 1)$



(b) Znečišťující rozdělení $\mathrm{N}(1, 1)$

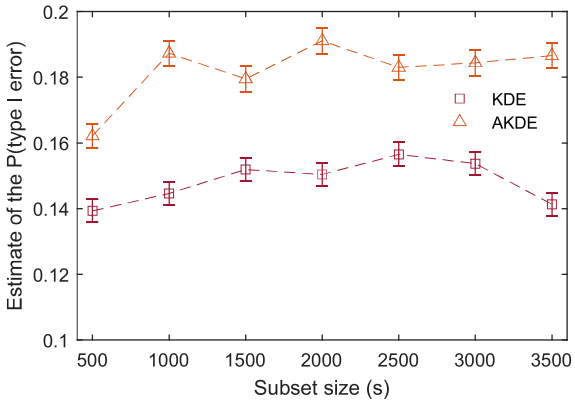


(c) Znečišťující rozdělení $\mathrm{N}(0.5, 1.5)$

Obrázek 3.5: Výsledky simulací pro odhad síly testu v závislosti na míře znečištění jednoho z testovaných souborů dat (portion of error samples).

nebo chybou samotného KDE odhadu skutečné hustoty pravděpodobnosti. V předchozích simulacích byla pro generování nevážených pozorování z WKDE, resp. WAKDE, použita metoda generování z náhodného výběru popsaná v podkapitole 2.3.2. Pro vyloučení chyby způsobené touto metodou generování byly provedeny simulace se stejným nastavením všech parametrů a rozdělení, ale získání nevážených pozorování bylo provedeno metodou generování z křivky 2.3.2.

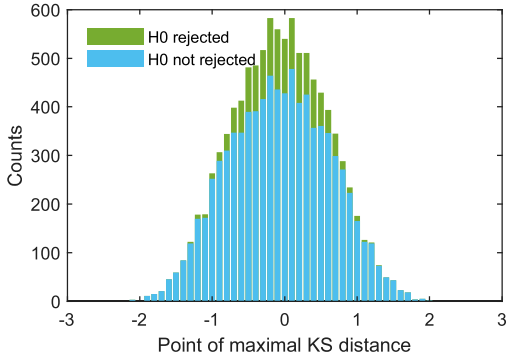
Testy s využitím jádrových odhadů a generováním z křivky byly provedeny pro pozorování generovaná z normálního rozdělení $\mathrm{N}(0, 1)$ a váhami z rozdělení $\mathrm{Beta}(4, 2)$. Odhady pravděpodobností chyby I. druhu jsou zobrazeny v grafu 3.6, ze kterého vidíme, že při generování přímo z křivky jádrového odhadu dostáváme velmi podobné výsledky, jako při generování s využitím testovaného souboru dat. Příčinou vysoké hodnoty pravděpodobnosti chyby I. druhu tedy není chyba v metodě generování z náhodného výběru.



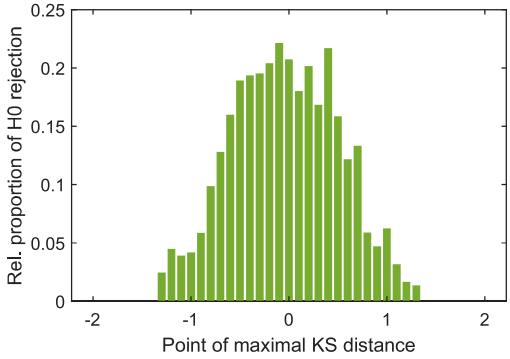
Obrázek 3.6: Odhady pravděpodobnosti chyby I. druhu pro testy s jádrovými odhady a generováním z křivky.

Příčinou vysoké hodnoty odhadu pravděpodobnosti chyby I. druhu je tedy samotná odlišnost WKDE, respektive WAKDE od skutečného rozdělení dat. Pro variantu testu s WKDE byla snaha zjistit, čím je způsobena vyšší míra zamítání H_0. Jako modelové rozdělení bylo opět použito normální rozdělení $\mathrm{N}(0, 1)$ pro hodnoty $\boldsymbol{x}$ a $\boldsymbol{y}$, používaná metoda volby jádra je pro data z normální rozdělení z hlediska chyby MISE (2.32) optimální. Pro váhy bylo použito rozdělení $\mathrm{Beta}(4, 2)$.

Testovací statistika použitého KS testu je založena na supremu rozdílu mezi empirickými distribučními funkcemi $\sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)|$ testovaných souborů dat. Opakovaným generováním dvojic vážených souborů o 1000 pozorováních a prováděním testů shodnosti jejich rozdělení bylo možné získat histogram 3.7, který zobrazuje, v jakých bodech x bylo nalezeno supremum vzdálenosti distribučních funkcí a kolikrát pro takový bod došlo k zamítnutí nebo nezamítnutí hypotézy H_0. Vedlejší obrázek 3.8 pak pro každý bin histogramu 3.7 znázorňuje, jaký byl podíl zamítnutí H_0 při nalezení suprema v bodě z daného binu k celkovému počtu simulací, ve kterých bylo supremum v daném binu nalezeno. Z tohoto grafu vidíme, že k častějšímu zamítání H_0 docházelo pro body suprema, které se vyskytovaly v okolí 0, problémem je tedy rozdíl ECDF nevážených dat (a tedy i rozdíl získaných WKDE) v hlavní části rozdělení, nikoliv ve chvostech, jak by se u jádrových odhadů dalo očekávat.



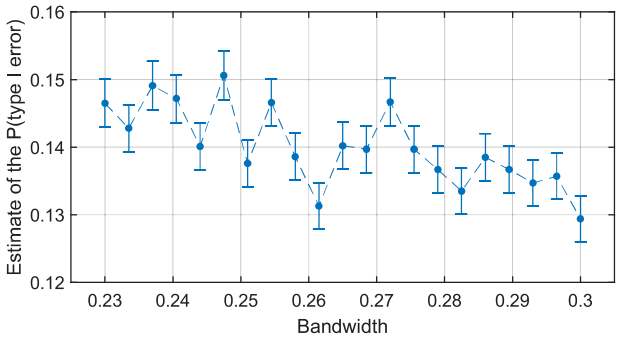
Obrázek 3.7: Histogram bodů, ve kterých bylo nalezeno supremum vzdálenosti distribučních funkcí při testování homogenity KS testem.



Obrázek 3.8: Relativní četnost zamítnutí H_0 pro bod suprema vzdálenosti dvou ECDF generovaných nevážených souborů.

Pro analýzu vztahu mezi parametrem šířky okna h a pravděpodobností chyby I. druhu pro testy s jádrovými odhady byla provedena také simulace, ve které byly hodnoty parametru h zvoleny pevně. Rozdělení pozorování a vah a velikost testovaných souborů zůstala stejná jako v předchozí simulaci, ale pro šířku okna byly pevně určeny hodnoty z rozsahu $0,23 - 0,3$, který odpovídá rozsahu hodnot h získaných pro danou velikost souborů dat a daná rozdělení při opakovaném výběru pozorování a výpočtu optimálního h dle vzorce (2.48).

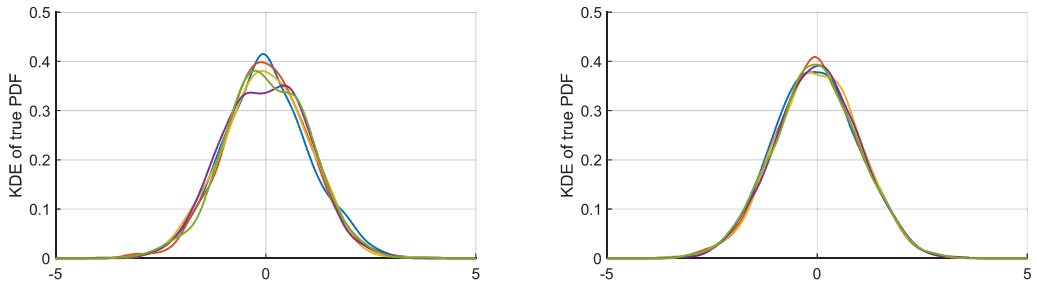
Graf 3.9 zobrazuje, jak se měnila pravděpodobnost chyby I. druhu pro různé hodnoty parametru h. Je zřetelné, že pravděpodobnost chyby I. druhu klesá s rostoucí šířkou okna. Tento jev lze vysvětlit tak, že s větší šířkou okna h dochází při jádrovém odhadu k většímu vyhlazování náhodné informace obsažené v datech. Pokud tedy konstruujeme dva WKDE a pro každý z nich používáme jinou množinu dat ze stejného rozdělení, pak pro větší hodnoty h dochází k většímu vyhlazení náhodných fluktuací v datech. Při testování dvou vážených souborů dat tedy nemá význam korigovat šířku okna dle dosažené hodnoty odhadu pravděpodobnosti chyby I. druhu.



Obrázek 3.9: Odhad pravděpodobnosti chyby I. druhu v závislosti na pevně zvolené hodnotě parametru h pro jádrový odhad. Hodnoty pozorování byly generovány z $\mathrm{N}(0,1)$, váhy z rozdělení $\mathrm{Beta}(4,2)$.

Tento jev také vysvětluje, proč na obrázku 3.3 pozorujeme vyšší pravděpodobnosti chyby I. druhu pro WAKDE. Adaptivní jádrový odhad má proměnlivou šířku okna h_j podle četnosti dalších pozorování v okolí bodu X_j, náhodné jevy vyskytující se v datech tedy vyhlazuje méně než WKDE se šířkou okna h stejnou jako pilotní šířka $\hat{f}_{pilot}$.

Obrázky 3.10 ilustrují, jak se mohou lišit jádrové odhady opakovaně konstruované na množinách o stejném počtu pozorování, které byly generované ze stejného rozdělení $\mathrm{N}(0,1)$. Hodnota parametru h byla určena automaticky dle (2.48). Skutečně lze pozorovat, že nejvíce viditelné rozdíly nastávají v hlavní části odhadnutých hustot, nikoliv ve chvostech. Analogické grafy pro další rozdělení jsou k dispozici v příloze A.3. Opět je z nich zřetelné, že jednotlivé křivky jádrových odhadů se vždy liší v hlavní části daného rozdělení, ve chvostech se viditelné rozdíly v podstatě nevyskytují.



(a) $s = 1000$ (b) $s = 2000$

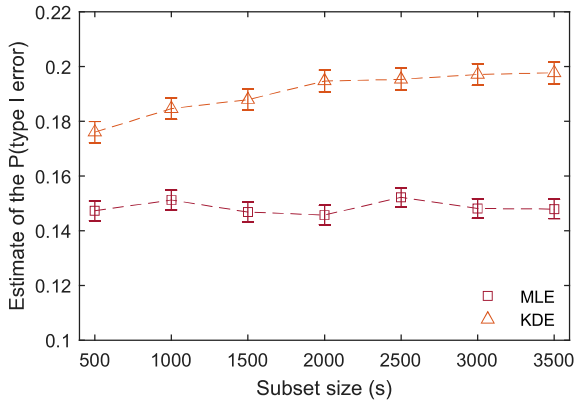
(c) $s = 3000$ (d) $s = 4000$

Obrázek 3.10: Každá z křivek v grafu odpovídá zobrazení KDE zkonstruovanému na jedné konkrétní realizaci náhodného výběru s hodnot z $\mathrm{N}(0,1)$. Pro každé s bylo provedeno 5 realizací náhodného výběru a konstrukce KDE.

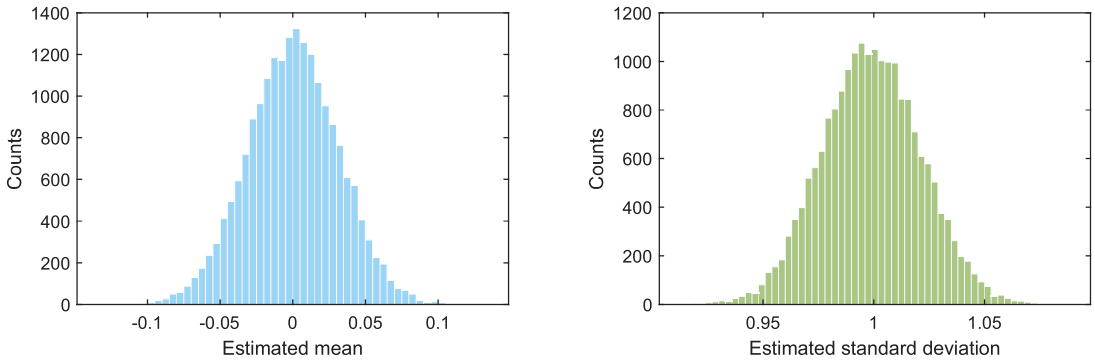
Pro získání dalšího vhledu do fungování testů homogenity s využitím jádrových odhadů byla provedena také simulace, při které byla testována shodnost rozdělení pro dva nevážené soubory dat generované z normálního rozdělení $\mathrm{N}(0,1)$. V první variantě testu byl opět použit KDE pro odhad hustoty pravděpodobnosti z dat a následné generování nové množiny pozorování. Ve druhé variantě testu byla předpokládána znalost normální rodiny rozdělení dat, ale parametry tohoto rozdělení (střední hodnota μ a směrodatná odchylka σ) byly z dat odhadnuty metodou maximální věrohodnosti (MLE).

Odhady pravděpodobnosti chyby I. druhu jsou zobrazeny v grafu 3.11. Přestože v tomto případě lze od parametrického odhadu hustoty očekávat větší přesnost než od neparametrického KDE, je pravděpodobnost chyby I. druhu pro variantu testu s využitím MLE odhadů parametrů stále výrazně vyšší než zvolená hladina významnosti $\alpha = 0,05$ a po-

měrně blízká hodnotám odhadu P(chyba I. druhu) získaným pro test s použitím KDE. Pro kontrolu MLE odhadů parametrů normálního rozdělení jsou na obrázku 3.12 vykresleny histogramy odhadů střední hodnoty $\hat{\mu}_{MLE}$ a směrodatné odchylky $\hat{\sigma}_{MLE}$.



Obrázek 3.11: Výsledky simulací testů homogenity s využitím neparametrických jádrových odhadů (KDE) nebo maximálně věrohodných odhadů parametrů rozdělení (MLE) pro dva soubory dat o s pozorováních z rozdělení $\mathrm{N}(0, 1)$.



(a) Odhady $\hat{\mu}_{MLE}$ (b) Odhady $\hat{\sigma}_{MLE}$

Obrázek 3.12: Histogramy MLE odhadů parametrů normálního rozdělení získané při simulacích testů homogenity pro nevážené soubory dat o $s = 1000$ pozorováních z rozdělení $\mathrm{N}(0, 1)$.

3.1.3 Testování homogenity vážených a nevážených dat

Úloha testů homogenity pro vážená data byla inspirována potřebou ověření shodnost rozdělení vážených simulací fyzikálních dat s neváženými reálnými měřeními v HEP. Pro úplnost byly tedy pro čtyři zkoumané varianty testu provedeny obdobné simulace, ve kterých tentokrát byla testována shodnost rozdělení váženého souboru dat $(\boldsymbol{x}, \boldsymbol{w})$ a neváženého souboru $\boldsymbol{y}$.

Protože celkově pro pozorování $\boldsymbol{x}$ po přiřazení vah $\boldsymbol{w}$ neznáme skutečné rozdělení, před samotnými simulacemi testů bylo nutné získat nevážená pozorování $\boldsymbol{y}$ se shodným

rozdělením, jako mají data $(\boldsymbol{x}, \boldsymbol{w})$. Nejprve byl generován velký počet (konkrétně 10 milionů) pozorování $\boldsymbol{x}_{pool}$ z vybraného rozdělení a k nim stejný počet nezávislých vah $\boldsymbol{w}_{pool}$ rozdělení Beta nebo z uniformního rozdělení. Na tomto velkém souboru dat byl následně zkonstruován WKDE, ze kterého byly generovány menší soubory nevážených pozorování $\boldsymbol{y}$, jejichž počet odpovídal součtu vah v datasetu $(\boldsymbol{x}, \boldsymbol{w})$ o s pozorováních, které byly náhodně vybírány z počáteční množiny 10 mil. vážených pozorování $(\boldsymbol{x}_{pool}, \boldsymbol{w}_{pool})$. Pro tyto menší soubory pozorování pak byla opakovaně zkoumána shodnost jejich rozdělení vybranou variantou KS testu. Ačkoliv předchozí výsledky poukázaly na chybu jádrového odhadu, v tomto případě je WKDE konstruován s využitím extrémně velkého počtu pozorování, který by měl zajistit dostatečně kvalitní odhad skutečného rozdělení [22]. Počet opakování testů byl opět $k = 10\,000$.

Odhady pravděpodobnosti chyby I. druhu pro testování váženého souboru s neváženým, kde hodnoty $\boldsymbol{x}$ byly generovány z různých rodin rozdělení a váhy $\boldsymbol{w}$ z rozdělení Beta$(4, 2)$ jsou zobrazeny v grafech 3.13. Testy s modifikovanou statistikou opět poskytují dobré výsledky, odhad pravděpodobnosti chyby I. druhu přibližně odpovídá zvolené hladině významnosti $\alpha = 0,05$. Pro testy využívající metodu re-arrangingu je odhadnutá pravděpodobnost chyby I. druhu zhruba 0,025, tedy vyšší než pro testování dvou vážených souborů, ale stále značně pod zvolenou hladinou významnosti $\alpha = 5$ %.

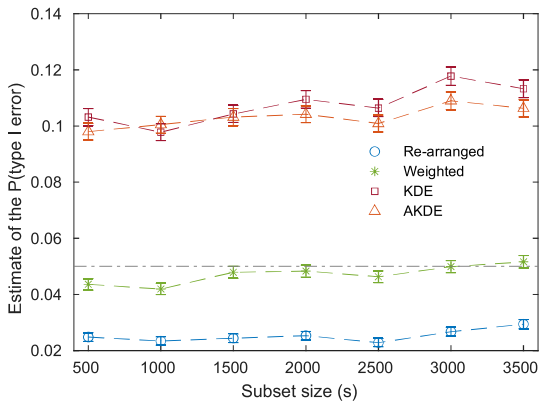
Testy používající jádrové odhady mají naopak nižší odhad pravděpodobnosti chyby I. druhu než při srovnávání dvou vážených souborů. Lze se tedy domnívat, že při testování dvou vážených souborů se v testu zkombinuje chyba vnesená oběma odhady a způsobí častější zamítnutí hypotézy H_0.

Zajímavý je výsledek pro pozorování $\boldsymbol{x}$ generovaná z lognormálního rozdělení, které má těžké chvosty. Při testování dvou vážených souborů byl při použití testu s WKDE, resp. WAKDE odhad P(chyby I. druhu) přibližně stejný, jako pro data z jiných rozdělení, test s WKDE vykazoval odhad chyby o trochu nižší než test s WAKDE. Pro testování váženého souboru s neváženým lze pozorovat, že odhad P(chyby I. druhu) pro test s WKDE je výrazně vyšší než pro test s WAKDE nebo při testování dvou vážených souborů. Zde se tedy výrazně projevila lepší kvalita WAKDE odhadu v těle i ve chvostech rozdělení.

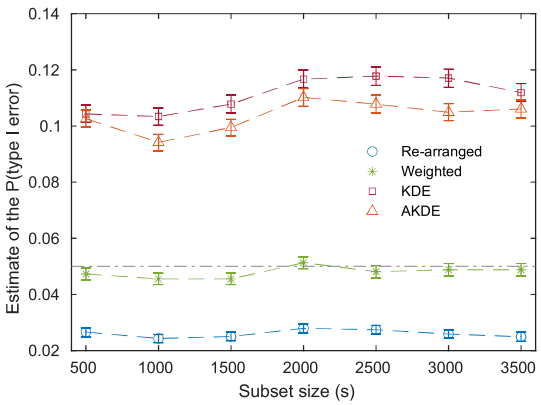
V případě testování dvou vážených souborů je zjevně WKDE značně odlišný od skutečného rozdělení, ale pro oba porovnávané datasety došlo u obou WKDE k podobné chybě, takže oba výsledné WKDE od sebe příliš odlišné nebyly. Proto byl odhad pravděpodobnosti chyby I. druhu na podobné hodnotě jako u ostatních rozdělení pozorování $\boldsymbol{x}$. V případě testování váženého souboru s neváženým byla transformace dat generováním z WKDE, resp. WAKDE použita pouze u vážených pozorování, chyba způsobená jádrovým odhadem se tedy projevila pouze u tohoto souboru a při porovnání s neváženým datasetem, který nebyl nijak transformován, pozorujeme častější zamítnutí H_0. Pro lognormální rozdělení s těžkými chvosty se tak ukázala i větší přesnost WAKDE díky úpravě šířky jádra h_j podle četnosti dalších pozorování v okolí hodnoty x_j.

3.1.4 Shrnutí

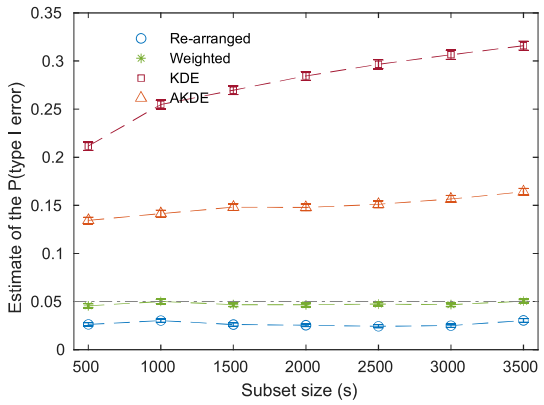
Přístup založený na odhadu hustoty pravděpodobnosti z dat metodou jádrových odhadů je v samotném principu úplně odlišný od testů s modifikovanou testovací statistikou i metody re-arranging. Při užití jádrových odhadů lze díky možnosti generování nevážených pozorování použít na takto získaná data klasický KS test, pro který je dokázáno



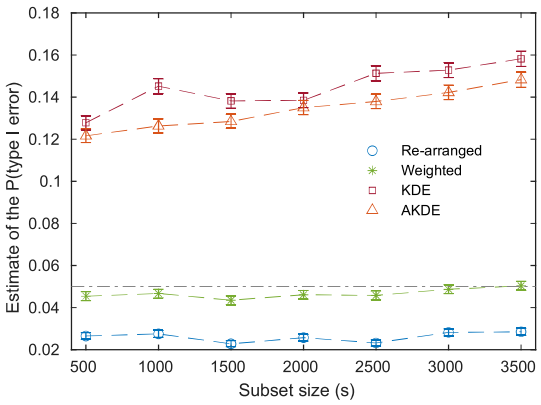
(a) Normální rozdělení



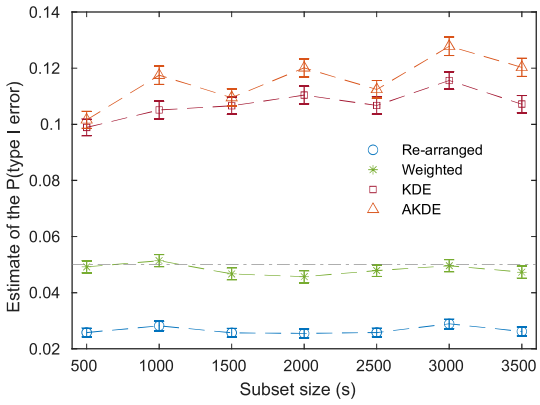
(b) Logistické rozdělení



(c) Lognormalní rozdělení



(d) Gamma rozdělení



(e) Weibullovo rozdělení

Obrázek 3.13: Odhad pravděpodobnosti chyby I. druhu pro testy váženého souboru o s pozorováních s neváženou množinou pozorování. Rozdělení pozorování $\boldsymbol{x}$ prvního souboru je uvedeno v popiscích grafů, rozdělení vah $\boldsymbol{w}$ bylo Beta$(4, 2)$.

asymptotické rozdělení testovací statistiky. Problémem ale je, že KS testem už nerozhodujeme o shodnosti rozdělení původních vážených dat, ale generovaných nevážených dat, jejichž rozdělení je určeno jádrovým odhadem a je tedy mírně odlišné od původního rozdělení vážených dat. Chyba vnesená jádrovým odhadem je dostatečně významná na to, aby ji KS test zachytil, proto je odhad pravděpodobnosti chyby I. druhu u těchto testů vyšší, než hladina významnosti α.

Jak lze vidět z grafu 3.11, stejný problém se projeví i v situaci, kdy testujeme homogenitu nevážených dat, u kterých známe rodinu rozdělení, z dat odhadneme pouze příslušné parametry, a následně z rozdělení s odhadnutými parametry generujeme novou množinu pozorování. Přestože takový odhad je v případě normální rodiny rozdělení přesnější než neparametrický KDE, tak odhadnutá P(chyby I. druhu) neklesá dostatečně. Princip testování vážených dat založený na konstrukci odhadu hustoty pravděpodobnosti a následném generování nových pozorování tedy není vhodné používat s asymptotickým rozdělením standardní testovací statistiky KS testu (2.12) bez vhodné korekce rozhodovacího kritéria pro zamítnutí hypotézy H_0.

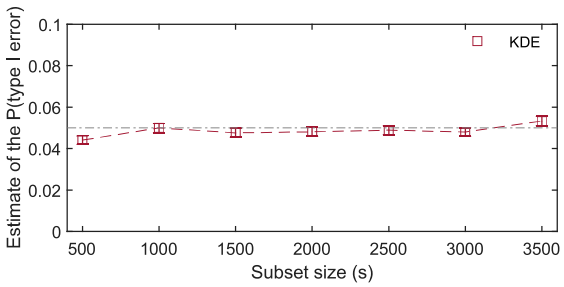
Simulace pro různé rodiny rozdělení ale ukázaly, že odhadnuté hodnoty pravděpodobnosti chyby I. druhu pro testy s jádrovými odhady byly velmi podobné pro různé rodiny rozdělení. Nabízí se tedy možnost stanovení nové modifikované kritické hranice pro p-hodnotu, pomocí které by se rozhodovalo o zamítnutí H_0. Pokud je možné z váženého histogramu nebo WEDF vážených dat přibližně určit rodinu rozdělení, lze provést numerickou simulaci testů s využitím jádrových odhadů pro tuto přibližnou rodinu rozdělení, analogicky k simulacím prezentovaným v této kapitole. Ze simulací je možné získat ECDF p-hodnot obdobně jako na obrázku 3.4. Pro alespoň přibližné naladění testu s jádrovými odhady na P(chyby I. druhu) $= \alpha$ pak kritickou hranici pro p-hodnotu stanovíme jako

$$p_{krit} = ECDF^{\leftarrow}_{p\text{-hodnota}}(\alpha) \tag{3.3}$$

a hypotézu H_0 zamítáme pokud p-hodnota $< p_{krit}$.

Tento postup byl pro ilustraci aplikován pro testování dvou souborů vážených dat s hodnotami pozorování $\boldsymbol{x}$ z N$(0, 1)$ a váhy $\boldsymbol{w}$ z Beta$(4, 2)$. Vzhledem k tomu, že WEDF takových dat je podobná ECDF normálního rozdělení N$(0, 1)$, byly provedeny simulace testů s jádrovými odhady pro normální rozdělení se zaznamenáním spočtených p-hodnot. Pro naladění testu na P(chyby I. druhu) $= 0,05$ byla z empirické distribuce p-hodnot určena kritická hranice pro zamítání H_0 na základě p-hodnoty testu $p_{krit} = 0,0092$. Výsledky opakovaného testování pro vážená data při kritické hranici p_{krit} jsou zobrazeny v grafu 3.14. Test se tímto způsobem pomocí simulace podařilo přibližné naladit na požadovanou pravděpodobnost chyby I. druhu.

Na základě získaných výsledků lze také doporučit použití testu homogenity s testovací statistikou modifikovanou pro vážená data (2.20) pomocí efektivní velikosti vzorku. Přestože pro tuto statistiku není odvozené její asymptotické rozdělení, numerické simulace prezentované v této kapitole ukázaly, že i při použití asymptotického rozdělení statistiky standardního KS testu je pro testování homogenity dvou vážených souborů dat přibližně dodrženo P(chyby I. druhu) $= \alpha$, kde α je zvolená hladina významnosti. Přestože odhady P(chyby I. druhu) byly prováděny pouze pro $\alpha = 0,05$, tak z grafů ECDF p-hodnot tohoto testu je patrné, že požadavek P(chyby I. druhu) $= \alpha$ bude přibližně splněn i při jiné volbě hladiny významnosti α.



Obrázek 3.14: Výsledky testů s KDE naladěnými na požadovanou P(chyba I. druhu) pomocí simulace.

Ve srovnání s předchozími výsledky analýzy testů homogenity s modifikovanou statistikou prezentovanými v [37, 38] bylo numerické ověření funkčnosti modifikovaných testů v této práci rozšířeno na více rodin rozdělení pozorování a na obecnější případ testování dvou vážených souborů dat.

Pro testy využívající metodu re-arranging bylo ukázáno, že přestože při jejich užití pravděpodobnost chyby I. druhu nepřekročí zvolenou hladinu významnosti α, tak jejich užití nelze doporučit z důvodu nízké hodnoty síly testu. Pro tuto variantu testování homogenity vážených dat není zatím přesně jasné, proč k tomuto jevu dochází. V publikaci [14] byla metoda re-arranging použita jako referenční pro ověření fungování testů homogenity s modifikovanou testovací statistikou. Výsledky uvedené v této práci prokázaly, že testy využívající re-arranging nelze považovat za vhodnou referenční metodu.

Kapitola 4

Generativní kompetitivní sítě

Modely hlubokého učení založené na neuronových sítích se úspěšně prosadily v mnoha aplikacích, obzvlášť pak pro klasifikační úlohy, které vyžadují mapování vysoce dimenzionální vstup na jednoduchý výstup [11]. Tím může být může například pravděpodobnost příslušnosti daného pozorování do jedné z předem určených tříd. Generativní kompetitivní sítě (anglicky generative adversarial networks, zkráceně GANs), představené roku 2014 [25], jsou generativním modelem využívajícím metody hlubokého učení k napodobení vícerozměrného rozdělení dat trénovací množiny. Přestože tento přístup neumožňuje získat přesnou formulaci odhadnutého rozdělení, lze s pomocí GAN generovat nové objekty, jejichž rozdělení je v ideálním případě shodné s rozdělením trénovacích dat.

Model GAN je tvořen dvěma neuronovými sítěmi. V případě použití GAN pro generování obrazových dat se navíc často jedná o konvoluční neuronové sítě [25]. Tato kapitola proto nejprve obsahuje stručné shrnutí základních informací o konvolučních neuronových sítích. Další část je věnována principu fungování GAN. Poslední sekce popisuje konkrétní model s názvem 3DGAN, který byl zmíněn již v kapitole 1 a který slouží ke generování 3-rozměrných simulací částice v kalorimetru ECAL, který byl navržen pro budoucí urychlovač částic CLIC v CERN [4].

V poslední kapitole 5 bude model 3DGAN použit pro ukázku implementace metody evaluace kvality GAN, která je založená na narozeninovém problému [10]. Předmětem této části práce nejsou úpravy architektury modelu 3DGAN ani modifikace učícího algoritmu. Proto jsou informace o neuronových sítích poskytnuté v této kapitole velmi stručné, pouze pro dodání základního kontextu.

4.1 Vícevrstvé neuronové sítě

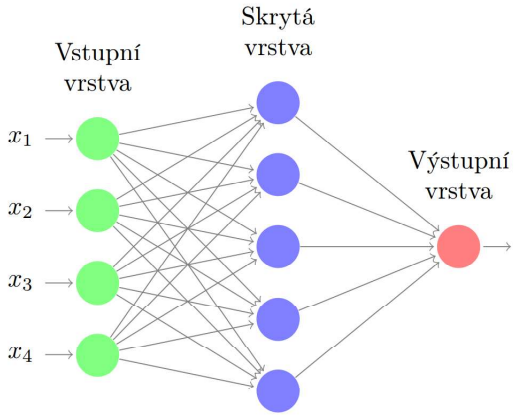
Vícevrstvý perceptron (anglicky multi-layer perceptron, MLP) představuje základní model strojového učení. S použitím trénovací množiny dat se učí aproximovat funkci $f : \mathbb{R}^m \to \mathbb{R}^o$, kde $m \in \mathbb{N}$ je dimenze vstupních dat a $o \in \mathbb{N}$ je dimenze výstupu.

MLP model sestává nejméně ze třech vrstev [3]. První je vstupní vrstva, následuje jedna nebo více skrytých vrstev a poslední je tzv. výstupní vrstva, jak znázorňuje schéma 4.1. Vstupní vrstva pouze předává vstupní příznaky $x_1, x_2, \ldots, x_m$ do neuronů v první skryté vrstvě. Každý neuron skryté vrstvy nejprve provede vážený součet svých vstupních příznaků pro váhy $w_i \in \mathbb{R}$, $\forall i \in \{1, \ldots, m\}$, na který následně aplikuje aktivační funkci

$g : \mathbb{R} \to \mathbb{R}$, neboli

$$y(\boldsymbol{x}, \boldsymbol{w}) = g\left(\sum_{i=1}^{m} w_i x_i\right), \tag{4.1}$$

kde $y(\boldsymbol{x}, \boldsymbol{w})$ výstupní hodnota neuronu pro vektor vstupních příznaků $\boldsymbol{x}$ a váhy $\boldsymbol{w}$. Výstup neuronu první skryté vrstvy je pak použit buď jako vstupní příznak pro neurony výstupní vrstvy (pokud se jedná pouze o třívrstvý MLP), nebo je vstupním příznakem pro neurony v další skryté vrstvě. Poslední (výstupní) vrstva MLP je také tvořena neurony, pro které jsou vstupními příznaky výstupy z poslední skryté vrstvy. Také neurony výstupní vrstvy provedou nejprve vážený součet svých vstupních příznaků, který ale následně transformují na výstupní hodnoty sítě.



Obrázek 4.1: Schéma třívrstvého MLP [28].

Aktivační funkce výstupní vrstvy je obvykle volena podle typu požadovaného výstupu. Pokud je cílem použití MLP binární klasifikace, pak je vhodnou aktivační funkcí ve výstupní vrstvě logistická sigmoidální funkce $g : \mathbb{R} \to [0, 1]$

$$g(t) = \frac{1}{1 + e^{-t}},$$

která mapuje hodnotu váženého součtu vstupních příznaků neuronu na interval $[0, 1]$. Výstupní hodnotu je pak možné interpretovat jako pravděpodobnost příslušnosti daného vstupního pozorování k jedné ze tříd.

Pokud je MLP použitý pro úlohu regrese, lze v poslední vrstvě použít jednoduchou lineární funkci $g : \mathbb{R} \to \mathbb{R}$

$$g(t) = t.$$

Další často používanou aktivační funkcí ve výstupní vrstvě je například ReLU (rectified linear unit) $g : \mathbb{R} \to \mathbb{R}_0^+$

$$g(t) = \max\{0, t\}, \tag{4.2}$$

která umožňuje mapování na nezáporné hodnoty. Ve specifických případech se užívá také funkce softmax $g : \mathbb{R}^l \to (0, 1)^l$

$$g_i(\boldsymbol{t}) = \frac{e^{t_i}}{\sum_{j=1}^{l} e^{t_j}} \quad \text{pro } i \in \{1, \ldots, l\}, \tag{4.3}$$

která je typická pro úlohy klasifikace do více tříd [24].

Trénováním MLP rozumíme opakované upravování vah sítě tak, aby pro trénovací množinu pozorování byla minimalizována ztrátová funkce MLP, která vyčísluje rozdíl mezi požadovaným výstupem a predikovanou hodnotou. Minimalizování ztrátové funkce je optimalizační úloha, kterou je možné řešit různými algoritmy, obvykle založenými na metodě gradientního sestupu (anglicky gradient descent, více například v [3, 24]).

Speciálním případem hlubokých neuronových sítí jsou konvoluční neuronové sítě (anglicky convolutional neural netwroks, CNNs). Jedná se o modely hlubokých neuronových sítí obsahující alespoň jednu vrstvu, ve které je na vstupní data aplikována operace konvoluce. V praxi se nejčastěji používají konvoluce na diskrétní 2D obrazová data. Konvoluci dvou funkcí $\boldsymbol{I},\ \boldsymbol{K} : \mathbb{Z}^2 \to \mathbb{R}^2$ proto definujeme jako

$$(\boldsymbol{I} * \boldsymbol{K})(i,j) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} \boldsymbol{X}(i-m, j-n)\boldsymbol{K}(m,n). \tag{4.4}$$

V praktické aplikaci $\boldsymbol{I}$ představuje obraz konečných rozměrů a $\boldsymbol{K}$ konvoluční jádro, které je rovněž konečných rozměrů. Nekonečná suma tedy přechází v sumu konečnou, určenou rozměry obrazu $\boldsymbol{I}$ a jádra $\boldsymbol{K}$. Volnými parametry konvoluční vrstvy nejsou v tomto případě váhy, ale prvky konvolučního jádra.

Definici diskrétní konvoluce lze pro $\boldsymbol{I},\ \boldsymbol{K} : \mathbb{Z}^3 \to \mathbb{R}^3$ rozšířit také na 3-rozměrná obrazová data předpisem

$$(\boldsymbol{I} * \boldsymbol{K})(i,j,k) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} \boldsymbol{X}(i-m, j-n, k-p)\boldsymbol{K}(m,n,p). \tag{4.5}$$

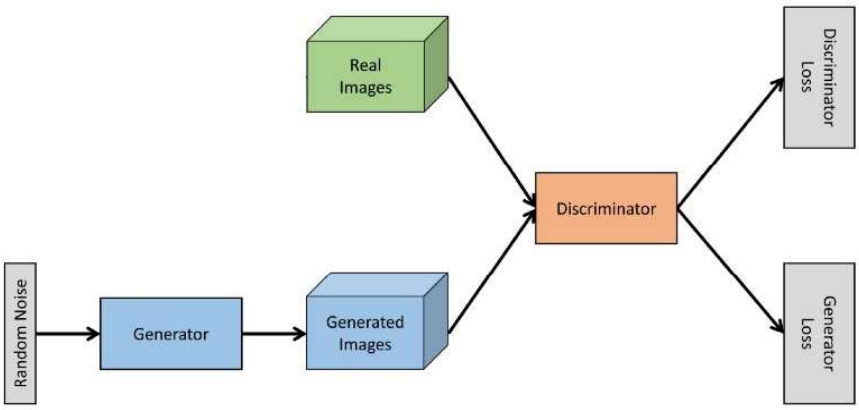
CNN našly své využití především v oblasti zpracování obrazových dat díky své schopnosti zohlednění informace z okolí právě zpracovávaného pixelu. Velikost tohoto okolí odpovídá rozměrům konvolučního jádra $\boldsymbol{K}$. Druhou výhodou konvoluční vrstvy je, že obsahuje menší počet parametrů, než tzv. plně propojená vrstva neuronové sítě, ve které jsou obyčejné neurony a každý z nich je propojený jak se všemi neurony v předchozí vrstvě, tak se všemi neurony vrstvy následující a ke každému z těchto spojů je přiřazena váha. V konvoluční vrstvě jsou parametry v konvolučním jádru, které je stejné pro všechny spoje.

4.2 Generativní kompetitivní sítě (GANs)

V obecném pojetí je generativní kompetitivní síť tvořena trénovací množinou dat a dvěma neuronovými sítěmi - generátorem G a diskriminátorem D. Vstupem do generátoru je vektor náhodných hodnot, obvykle generovaný z normálního rozdělení. Tento vstup se generátor postupně učí mapovat na vícerozměrný výstup podobný pozorováním v trénovací množině. V podstatě se tedy generátor učí rozdělení dat trénovací množiny. Diskriminátor D naopak mapuje trénovací data nebo výstupy z generátoru na hodnotu z intervalu $[0, 1]$, která odpovídá pravděpodobnosti, že vstupní objekt pochází ze stejného rozdělení jako trénovací data. Diskriminátor tedy slouží k rozlišení trénovacích pozorování a pozorování produkovaných generátorem G.

Při trénování modelu GAN spolu generátor G a diskriminátor D soupeří. Diskriminátor D je pomocí trénovacích dat a pozorování získaných z G trénován tak, aby dokázal tyto dvě skupiny dat co nejlépe odlišit. Naopak cílem generátoru G je naučit se z trénovacích

dat jejich rozdělení a maximalizovat pravděpodobnost, že se diskriminátor D pro pozorování pocházející z G při klasifikaci splete. V podstatě se tedy jedná o tzv. minimaxní hru dvou hráčů, ve které generátor G vyhraje, když diskriminátor D pro libovolné vstupní pozorování vrátí výstupní hodnotu $\frac{1}{2}$. Základní struktura GAN je ilustrována schématem 4.2.



Obrázek 4.2: Schéma struktury GAN. Na vstupu generátoru (generator) je vektor náhodných hodnot (random noise). Výstupem generátoru jsou syntetická data (generated images). Diskriminátor (discriminator) slouží k rozlišení trénovacích dat (real images) od syntetických dat z generátoru. Cílem diskriminátoru i generátoru je minimalizovat příslušnou ztrátovou funkci (discriminator loss a generator loss) [31].

Pro formálnější zápis principu GAN popsaného výše předpokládejme, že náhodný vektor $\boldsymbol{Z}$ na vstupu generátoru pochází z rozdělení $\mathcal{D}_{\boldsymbol{Z}}$ na $\mathbb{R}^l$, obvykle se jedná o l-rozměrné normální rozdělení $\mathrm{N}_l(0, \mathbb{I}_l)$. Dále označme $\{G_u \,|\, u \in \mathcal{U}\}$ třídu generátorů, kde G_u je funkce $G_u : \mathbb{R}^l \to \mathbb{R}$, u je vektor parametrů funkce G_u a $\mathcal{U} \in \mathbb{R}^p$ označuje množinu možných vektorů parametrů. Generátor G_u pak definuje rozdělení $\mathcal{D}_{G_u}$ na $\mathbb{R}^d$ tak, že pokud je $\boldsymbol{Z} \sim \mathcal{D}_{\boldsymbol{Z}}$, pak řekneme, že náhodná veličina $\boldsymbol{X} = G_u(\boldsymbol{Z})$ má rozdělení $\mathcal{D}_{G_u}$. Dále označíme $\{D_v \,|\, v \in \mathcal{V}\}$ třídu diskriminátorů, kde D_v je funkce $D_v : \mathbb{R}^d \to [0, 1]$, v je vektor parametrů diskriminátoru a $\mathcal{V} \in \mathbb{R}^q$ je množina možných vektorů parametrů v. Skutečné rozdělení trénovacích dat označíme $\mathcal{D}_{real}$.

Při trénování diskriminátoru se snažíme nalézt takové parametry v, aby $D_v(\boldsymbol{x}) = 1$, pokud pozorování $\boldsymbol{x}$ pochází z rozdělení $\mathcal{D}_{real}$, a $D_v(\boldsymbol{x}) = 0$, pokud $\boldsymbol{x}$ bylo generováno z $\mathcal{D}_{G_u}$. Chceme tedy nalézt

$$\arg\max_{v \in \mathcal{V}} V(G_u, D_v) = \arg\max_{v \in \mathcal{V}} \left\{ \mathrm{E}_{\boldsymbol{X} \sim \mathcal{D}_{real}} \left[\phi(D_v(\boldsymbol{X})) \right] + \mathrm{E}_{\boldsymbol{X} \sim \mathcal{D}_{G_u}} \left[\phi(1 - D_v(\boldsymbol{X})) \right] \right\}, \quad (4.6)$$

kde $\phi : [0, 1] \to \mathbb{R}$ je tzv. měřící funkce. Trénováním generátoru zase chceme pro pevně daný diskriminátor D_v nalézt parametry u tak, abychom maximalizovali chybu diskriminátoru pro syntetická pozorování z generátoru, hledáme tedy

$$\arg\min_{u \in \mathcal{U}} \mathrm{E}_{\boldsymbol{X} \sim \mathcal{D}_{G_u}} \left[\phi(1 - D_v(\boldsymbol{X})) \right]. \quad (4.7)$$

Trénování modelu GAN spočívá v minimaxní optimalizaci účelové funkce $V(G_u, D_v)$

$$\min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} V(G_u, D_v) = \min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} \left\{ \mathrm{E}_{\boldsymbol{X}\sim\mathcal{D}_{real}} \left[\phi(D_v(X))\right] + \mathrm{E}_{\boldsymbol{X}\sim\mathcal{D}_{G_u}} \left[\phi(1 - D_v(X))\right] \right\}. \tag{4.8}$$

Nechť je měřící funkce ϕ monotónní a konkávní na intervalu $[0, 1]$. Pak platí, že pokud $\mathcal{D}_{real} = \mathcal{D}_{G_u}$, tak optimální strategií pro diskriminátor D_v je přiřadit jakémukoliv pozorování hodnotu $\frac{1}{2}$ a optimální hodnota účelové funkce je $2 \cdot \phi(\frac{1}{2})$.

Formulace optimalizační úlohy (4.8) předpokládá znalost rozdělení $\mathcal{D}_{real}$ a $\mathcal{D}_{G_u}$. Tato rozdělení ale v praxi neznáme, k dispozici máme pouze trénovací množinu dat a množinu pozorování získaných z generátoru. Střední hodnoty jsou proto při výpočtech nahrazeny aritmetickými průměry jako odhady skutečných středních hodnot.

V původním článku [25], ve kterém byly GANs poprvé představeny, je měřící funkcí $\phi(t) = \log t$. Tato volba je v aplikacích GAN velmi častá, ale může způsobovat problémy při trénování, protože $\log t \xrightarrow{t\to 0} -\infty$. Druhou často používanou měřící funkcí je identita $\phi(t) = t$, která byla poprvé použita v modelu Wasserstein GAN (zkráceně WGAN) [8].

4.2.1 Interpretace optimalizační úlohy

Minimaxní optimalizaci účelové funkce (4.8) lze teoreticky interpretovat jako úlohu minimalizace vzdálenosti mezi distribucemi $\mathcal{D}_{real}$ a $\mathcal{D}_{G_u}$. Předpokládejme, že je k dispozici nekonečně mnoho trénovacích dat a nekonečně mnoho syntetických pozorování z generátoru. Dále předpokládejme, že diskriminátor D může být libovolná neuronová síť (není omezen počet ani hodnoty jejích parametrů). Pak je pro pevně daný generátor G teoreticky možné získat optimální diskriminátor D^* minimalizující svou ztrátovou funkci

$$D^* = \arg\max_D V(G, D) = \arg\max_D \left\{ \mathrm{E}_{\boldsymbol{X}\sim\mathcal{D}_{real}} \left[\phi(D(\boldsymbol{X}))\right] + \mathrm{E}_{\boldsymbol{X}\sim\mathcal{D}_G} \left[\phi(1 - D(\boldsymbol{X}))\right] \right\}. \tag{4.9}$$

Je-li měřící funkcí $\phi(t) = \log t$, pak optimální diskriminátor D^* má podobu

$$D^*(\boldsymbol{x}) = \frac{p_{real}(\boldsymbol{x})}{p_{real}(\boldsymbol{x}) + p_G(\boldsymbol{x})}, \tag{4.10}$$

kde p_{real} a p_G jsou hustoty pravděpodobností příslušné k $\mathcal{D}_{real}$ a $\mathcal{D}_G$. Minimalizace chyby generátoru $\min_G \mathrm{E}_{\boldsymbol{x}\sim\mathcal{D}_G} \left[\log(1 - D^*(\boldsymbol{x}))\right]$ pak přechází na úlohu minimalizace Jensenovy-Shannonovy divergence rozdělení $\mathcal{D}_{real}$ a $\mathcal{D}_G$ [25]

$$\min_G d_{JS}(\mathcal{D}_{real}, \mathcal{D}_G) = \min_G \frac{1}{2} \left[d_{KL}\left(\mathcal{D}_{real}, \frac{\mathcal{D}_{real} + \mathcal{D}_G}{2}\right) + d_{KL}\left(\mathcal{D}_G, \frac{\mathcal{D}_{real} + \mathcal{D}_G}{2}\right) \right], \tag{4.11}$$

kde d_{KL} označuje Kullbackovu-Leiblerovu divergenci definovanou pro rozdělení $\mathcal{D}_1$, $\mathcal{D}_2$ na $\mathbb{R}^d$ vztahem $d_{KL}(\mathcal{D}_1, \mathcal{D}_2) = \int_{\mathbb{R}^d} \mathcal{D}_1(\boldsymbol{x}) \log\left(\frac{\mathcal{D}_1(\boldsymbol{x})}{\mathcal{D}_2(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$.

Uvažujme nyní měřící funkci $\phi(t) = t$ a funkci diskriminátoru D pouze z užší množiny 1-Lipschitzovských funkcí[1]. Dále předpokládejme, že D^* je mezi 1-Lipschitzovskými funkcemi optimální z hlediska minimalizace ztrátové funkce diskriminátoru. Pak úloha

[1]Funkce $f : \mathbb{R}^d \to \mathbb{R}$ je K-Lipschitzovská právě tehdy, když $(\forall \boldsymbol{x}_1, \boldsymbol{x}_2)(|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \leq K\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|)$.

minimalizace chyby generátoru $\min_G \mathrm{E}_{\boldsymbol{x}\sim\mathcal{D}_G}\left[1 - D^*(\boldsymbol{x})\right]$ přechází na úlohu minimalizace Wassersteinovy vzdálenosti [8]

$$\min_G d_W(\mathcal{D}_{real}, \mathcal{D}_G) = \min_G \sup_{f \text{ 1-Lipschitz}} \left| \mathrm{E}_{\boldsymbol{X}\sim\mathcal{D}_{real}}\left[f(\boldsymbol{X})\right] - \mathrm{E}_{\boldsymbol{X}\sim\mathcal{D}_G}\left[f(\boldsymbol{X})\right] \right|. \tag{4.12}$$

Podotkněme, že výše uvedené vztahy jsou pouze teoretické, jejich předpoklady neomezenosti kapacity diskriminátoru a nekonečného množství trénovacích dat nelze v praxi splnit.

4.2.2 Trénování GAN

Trénování GAN lze rozdělit do dvou fází - trénování diskriminátoru a trénování generátoru - které se opakovaně střídají. Nejprve v k iteracích probíhá trénování diskriminátoru:

1. generujeme m latentních vektorů (náhodný šum) $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(m)}$ z rozdělení $\mathcal{D}_Z$, která použijeme jako vstupy do generátoru k získání m syntetických pozorování z $\mathcal{D}_G$,

2. vybereme m pozorování $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}$ z trénovací množiny dat,

3. provedeme aktualizaci parametrů diskriminátoru D_v přičtením stochastického gradientu
$$\nabla_v \frac{1}{m} \sum_{i=1}^{m} \left[\phi\left(D_v\left(\boldsymbol{x}^{(i)}\right)\right) + \phi\left(1 - D_v\left(G_u\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]. \tag{4.13}$$

Po k iteracích trénování diskriminátoru proběhne zpravidla pouze jedna iterace trénování generátoru:

1. generujeme m latentních vektorů $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(m)}$ z rozdělení $\mathcal{D}_Z$,

2. upravíme parametry generátoru G odečtením stochastického gradientu
$$\nabla_u \frac{1}{m} \sum_{i=1}^{m} \phi\left(1 - D_v\left(G_u\left(\boldsymbol{z}^{(i)}\right)\right)\right). \tag{4.14}$$

Pro aktualizaci parametrů generátoru a diskriminátoru může být využita libovolná učící metoda založená na gradientu ztrátové funkce [25].

Předpokládejme opět, že máme neomezený počet trénovacích pozorování. Dále předpokládejme, že diskriminátor D a generátor G mají oba neomezenou kapacitu, jinými slovy že mohou mít libovolný počet parametrů. Pak algoritmus popsaný výše zajistí nalezení optimálních modelů G a D a po skončení trénování bude platit $\mathcal{D}_{real} = \mathcal{D}_G$ [25]. Přestože toto tvrzení skutečně platí, je opět závislé na v praxi nesplnitelných předpokladech. V [9] bylo naopak pro GAN teoreticky odvozeno, že při omezeném počtu parametrů sítí a konečném počtu trénovacích dat může trénování skončit výhrou generátoru, i když rozdělení $\mathcal{D}_{real}$ a $\mathcal{D}_G$ jsou stále rozdílná.

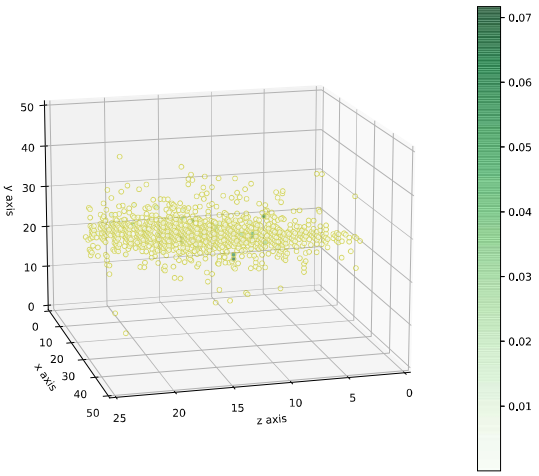
4.3 Model 3DGAN

Model 3DGAN vyvíjený výzkumnou skupinou CERN openlab byl již krátce popsán v kapitole 1. V této sekci popíšeme tento model trochu blíže, abychom v poslední kapitole 5 mohli ilustrovat použití metody pro odhad supportu založené na paradoxu narozenin právě pro model 3DGAN.

4.3.1 Data

Generativní model 3DGAN je vyvíjen s cílem získat rychlou metodu generování simulací interakcí elementárních částic v kalorimetru. Pro 3DGAN se konkrétně jedná o simulace elektromagnetického kalorimetru s vysokou granularitou, který byl navržen v rámci studie pro budoucí lineární urychlovač částic CLIC.

Trénovací množina dat pro 3DGAN je tvořena MC simulacemi interakcí elektronu s materiálem kalorimetru, které byly získány použitím softwarového nástroje Geant4 [6]. Velikost trénovací množiny pro 3DGAN čítá 400 000 pozorování. Simulace elektronů byly podmíněny hodnotou jejich počáteční energie E_p, která se pohybuje v rozsahu $5-200$ GeV. Zároveň byly výsledky simulací elektronů podmíněny také úhlem θ, pod kterým elektrony vstupují do kalorimetru a který se pohybuje v rozsahu $60° - 120°$ vzhledem k čelní stěně kalorimetru. Před použitím MC simulací pro trénování modelu 3DGAN byly všechny snímky předzpracovány, došlo k vycentrování těžiště události na snímku a jeho oříznutí na rozměry $51 \times 51 \times 25$. Obrázek 4.3 je ilustrací 3D snímku interakce elektronu v kalorimetru po vycentrování snímku a upravení rozměrů.

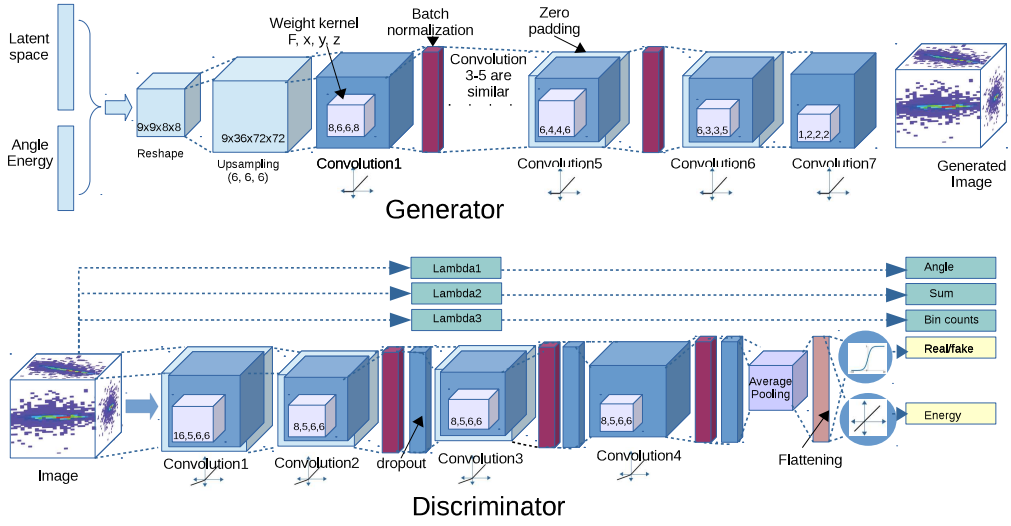


Obrázek 4.3: Příklad MC simulace elektronu v EM kalorimetru. Barva jednotlivých bodů odpovídá energii vyzářené částicemi v daném místě kalorimetru v jednotkách GeV.

4.3.2 Architektura 3DGAN

Schéma architektury generátoru a diskriminátoru modelu 3DGAN je znázorněno na obrázku 4.4. Vstupními daty generátoru je 254 hodnot generovaných z normálního rozdělení $\mathrm{N}(0, 1)$, navíc je k nim ale přidána také informace o počáteční energii částice E_p a úhlu θ, pod kterým částice vstupuje do kalorimetru. Generátor se tedy z trénovacích dat učí

vícerozměrné rozdělení, které je navíc podmíněné dvěma vstupními parametry, primární energií a úhlem.



Obrázek 4.4: Architektura modelu 3DGAN [16].

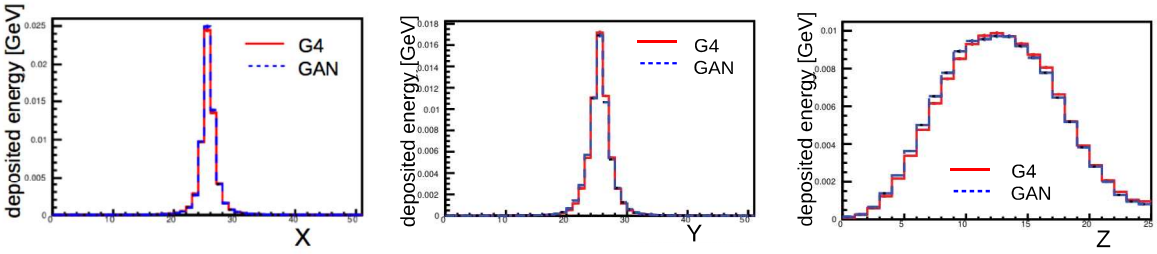
Pro vstupní hodnoty generátoru je nejprve proveden tzv. up-sampling, neboli opakované kopírování náhodného vektoru, dokud nezískáme objekt požadovaných rozměrů. Po up-samplingu následuje série 7 konvolučních vrstev, ve kterých jsou používány 3-rozměrné konvoluční masky. Výstupem z generátoru je 3-rozměrný snímek se stejnými rozměry, jaké mají MC simulace tvořící trénovací data.

Diskriminátor modelu 3DGAN produkuje dvojí výstup. První výstupní hodnota odpovídá pravděpodobnosti, že vstupní snímek pocházel ze stejného rozdělení jako trénovací data, a je získán užitím aktivační funkce sigmoid ve výstupní vrstvě sítě. Druhou výstupní hodnotou je odhad primární energie E_p částice na vstupním snímku, který je získán použitím lineární aktivační funkce v druhém z neuronů výstupní vrstvy diskriminátoru.

Trénování 3DGAN probíhalo nejprve na MC simulacích elektronů s počáteční energií v rozsahu $100-200$ GeV, aby v první fázi trénováni měla data menší variabilitu. Ve druhé fázi bylo trénovaní modelu rozšířeno na celý rozsah hodnot primární energie $2-500$ GeV.

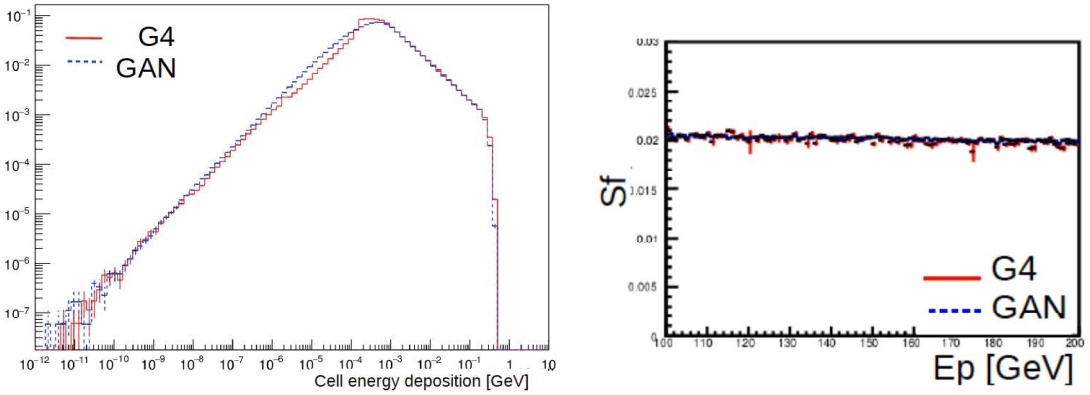
4.3.3 Validace modelu 3DGAN

Model 3DGAN byl autory v [16] validován srovnáním generovaných snímků s Monte Carlo simulacemi z programu Geant4, na kterých byl 3DGAN trénován. První validační metodou bylo porovnání množství energie vyzářené ve směrech hlavních os x, y a z, takzvaných *shower shapes*. Výsledné grafy průměrných hodnot vyzářené energie v závislosti na poloze vzhledem k hlavním osám kalorimetru jsou pro elektrony vstupující do kalorimetru kolmo ($\theta = 90°$) zobrazeny v grafech 4.5. Z těch je vidět velmi dobrá shoda MC simulací (označeno G4) se snímky generovanými z 3DGAN. Obdobně uspokojivé bylo toto srovnání také pro jiné hodnoty úhlu θ [16].



Obrázek 4.5: Porovnání distribuce energie podél hlavních os x, y, z pro MC simulace (G4) a 3DGAN simulace (GAN) částic vstupujících do kalorimetru kolmo k čelní stěně [16].

Hodnoty energií zaznamenaných jednotlivými pixely mají rozpětí několika řádů, proto byly MC simulace a 3DGAN snímky porovnány také z hlediska pravděpodobnosti výskytu pixelů o různých intenzitách. Toto srovnání je zobrazeno v grafu 4.6a. Opět vidíme, že histogram odpovídající snímkům z 3DGAN dobře kopíruje křivku pro MC simulace, některé nerovnosti v histogramu pro MC simulace jsou ale modelem 3DGAN zjevně vyhlazeny.



(a) Distribuce energie (b) Sampling fraction

Obrázek 4.6: (a) Histogram výskytu pixelů o intenzitě odpovídající hodnotě na ose x pro MC simulace a 3DGAN simulace. (b) Srovnání průměrných hodnot veličiny sampling fraction pro simulace částic o počáteční energii odpovídající hodnotě na ose x [16].

Důležitou charakteristikou záznamu elektronu v kalorimetru je také tzv. *sampling fraction*, neboli podíl energie vyzářené v kalorimetru a primární energie E_p. Pravý graf 4.6b proto zobrazuje srovnání průměrných hodnot sampling fraction (Sf) MC simulací a 3DGAN simulací v závislosti na počáteční energii E_p. Získaný graf opět ukazuje dobrou shodu mezi MC simulacemi a daty získanými z modelu 3DGAN.

Srovnání všech výše popsaných charakteristik naznačuje, že 3DGAN je schopen poskytnout kvalitní simulace, které jsou v poměrně dobré shodě s trénovacími MC simulacemi. Pro výpočet těchto charakteristik je ale informace obsažená v 3D snímcích agregována do jedné hodnoty, například pro sampling fraction, nebo vektoru hodnot v případě shower shapes. Pro simulace z generativních algoritmů v HEP v současnosti chybí jednotný přístup k validaci GAN simulací, který je nezbytný pro budoucí použití generativních modelů k vědeckým simulacím.

V následující kapitole popíšeme metodu pro validaci kvality dat simulovaných z GAN, která byla použita pro generativní modely napodobující reálné fotografie. Pokusíme se tutéž metodu přizpůsobit simulacím kalorimetru a aplikovat ji na model 3DGAN.

Kapitola 5

Odhad supportu modelu 3DGAN

Generativní modely v mnoha aplikacích ukázaly, že mají schopnost naučit se z trénovacích dat složitá, mnoharozměrná rozdělení a produkovat věrohodná syntetická data. Především jejich využití pro generování realistických napodobenin fotografií a dalších obrazových dat je úspěšné ([9], [33]). Teoretická analýza provedená v [9] nicméně ukázala, že i když se při trénování generativního modelu přiblížíme optimální hodnotě účelové funkce, není zaručeno, že naučené rozdělení je opravdu blízké skutečnému rozdělení trénovacích dat. Schopnost generátoru naučit se skutečnou distribuci $\mathcal{D}_{real}$ totiž závisí nejen na jeho vlastní kapacitě, ale také na kvalitě a rozsahu trénovacích dat a kapacitě diskriminátoru.

V [10] byl navržen postup pro získání odhadu supportu generativního modelu, který je založený na principu problému narozenin. Supportem GAN je v tomto kontextu myšlen support rozdělení, které se v procesu trénování generátor naučil z trénovacích dat. Aplikací této metody na vybrané modely GAN generující obrazová data bylo ukázáno, že i pro dobře etablované generativní modely natrénované na známých datasetech (například CIFAR-10 [27], CelebA [45]) je support naučeného rozdělení výrazně menší, než support rozdělení trénovacích dat.

Tato kapitola se nejprve věnuje popisu metody odhadu supportu generativního modelu založené na principu problému narozenin. Následuje specifikace způsobu použití této metody pro analýzu supportu modelu 3DGAN. Kapitola je zakončena aplikací metody odhadu supportu pro model 3DGAN a zhodnocením získaných výsledků.

5.1 Narozeninový problém

Skupina pouze 23 lidí stačí k tomu, aby pravděpodobnost, že dva lidé z této skupiny mají narozeniny ve stejný den v roce, byla 50 %. To je odpověď na otázku takzvaného *narozeninového problému*, který se poprvé objevil ve 20. letech 20. století [15]. Posléze začala být zkoumána také zobecněná varianta narozeninového problému, ve které uvažujeme obecný rok o $d \in \mathbb{N}$ dnech a opět se ptáme, jaký nejmenší počet osob n_d musí být v jedné místnosti, aby pravděpodobnost, že alespoň dva lidé mají narozeniny ve stejný den v roce, byla větší nebo rovna 50 %.

Pro získání n_d řešení zobecněného problému narozenin potřebujeme nalézt nejmenší $k \in \mathbb{N}$ takové, které splňuje

$$P(k) = \left(1 - \frac{1}{d}\right)\left(1 - \frac{2}{d}\right)\cdots\left(1 - \frac{k-1}{d}\right) \leq \frac{1}{2}, \tag{5.1}$$

kde $P(k)$ je vyjádření pravděpodobnosti, že ve skupině k lidí nemají žádné dvě osoby narozeniny ve stejný den z předpokladu nezávislosti dat narození těchto osob. Použitím následujících nerovností

$$P(k) \geq 1 - \left(\frac{1}{d} + \frac{2}{d} + \cdots + \frac{k-1}{d}\right) = 1 - \frac{k^2 - k}{2d}$$
$$P(k) \leq \exp\left(-\frac{1}{d}\right) \exp\left(-\frac{2}{d}\right) \cdots \exp\left(-\frac{k-1}{d}\right) = \exp\left(-\frac{k^2 - k}{d}\right)$$

je možné pro hodnotu n_d získat horní a dolní mez

$$\sqrt{d + \frac{1}{4}} + \frac{1}{2} \leq n_d < \sqrt{2d \log 2 + \frac{1}{4}} + \frac{3}{2}. \tag{5.2}$$

Odbočme nyní na chvíli od problému narozenin a uvažujme nyní posloupnost $A \subset \mathbb{N}$, jejíž prvky splňují $a_1 < a_2 < \cdots$. Pak pro množinu A definujeme *asymptotickou hustotu*

$$\delta_(A) = \lim_{n \to +\infty} \frac{A(n)}{n}, \text{ kde } A(n) = |\{a \in A \,|\, a \leq n\}|, \tag{5.3}$$

pokud tato limita existuje. Přirozená hustota je tedy určitým vyjádřením „velikosti“ podmnožiny přirozených čísel.

Od doby vyslovení problému narozenin byly odvozovány stále přesnější odhady řešení n_d. Teprve před přibližně 10 lety se podařilo ukázat řešení

$$n_d = \lceil \sqrt{2d \log 2} \rceil, \tag{5.4}$$

platné pro množinu přirozených čísel d s asymptotickou hustotou $\approx 0,731$ [15]. Stejnému autorovi se v roce 2012 podařilo vzorec pro výpočet n_d ještě zpřesnit na

$$n_d = \lceil \sqrt{2d \log 2} + \frac{3 - 2 \log 2}{6} \rceil \tag{5.5}$$

a dokázat, že platí pro množinu přirozených čísel d s asymptotickou hustotou 1 [15].

Záměrem této kapitoly je využití principu narozeninového problému pro odhadnutí velikosti supportu generativní sítě. Při této úloze si vystačíme s hrubou aproximací $n_d \approx d$. Předpokládejme na chvíli, že neznáme počet dní v roce d, ale naopak víme, jaký je nejmenší počet osob n_d, který bude stále ještě splňovat podmínku

$$P(\textit{alespoň dvě osoby mají narozeniny ve stejný den}) \geq \frac{1}{2}. \tag{5.6}$$

Pak díky znalosti přibližného vztahu $n_d \approx d$ můžeme hodnotou n_d^2 odhadnout počet unikátních dní v roce.

5.2 Adaptace narozeninového problému pro 3DGAN

Při aplikaci principu obráceného narozeninového problému pro odhad supportu generativního modelu budeme nejprve generovat z GAN množinu s pozorování. Následně

zkontrolujeme, zda se v této množině nachází alespoň jedna dvojice stejných objektů, neboli duplikátů. Dostatečným počtem opakování tohoto postupu je pak možné získat odhad pravděpodobnosti, že v množině s objektů se vyskytuje alespoň jedna dvojice shodných výstupů z GAN. Odhad velikosti supportu GAN pak dle poslední poznámky v předchozí sekci získáme jako s^{*2}, kde s^* označuje nejmenší velikost množiny dat z GAN, pro kterou je odhadnutá pravděpodobnost nalezení dvojice duplikátů ≥ 50 %. Úplně stejný postup lze aplikovat i na množinu trénovacích dat a získat tak odhad supportu skutečného rozdělení.

Zádrhelem tohoto přístupu je, jak pro obrazová data definovat duplikátní pozorování. V původní úloze problému narozenin bychom za duplikátní objekty označili ty osoby, které mají narozeniny ve stejný den. Takové označení je přímočaré a nekonfliktní, protože možných dnů narození je konečný počet a řídí se diskrétním rozdělením. Naproti tomu obrazová data jsou zpravidla tvořena velkým množstvím pixelů, kterým náleží nezáporné reálné číslo reprezentující jejich intenzitu. V případě obrázků o N pixelech se tedy jedná o data ze spojitého rozdělení na $\left(\mathbb{R}_0^+\right)^N$, takže pravděpodobnost nalezení přesně stejných objektů je nulová.

Uvažujme tedy obrazová data o N pixelech. V takovém případě můžeme za duplikáty považovat ty objekty, které jsou si „dostatečně podobné“ vzhledem k předem stanoveným kritériím podobnosti, tedy funkcím zobrazujícím $\left(\mathbb{R}_0^+\right)^N \times \left(\mathbb{R}_0^+\right)^N \to \mathbb{R}$. Od vhodného kritéria podobnosti ϱ očekáváme, že bude omezené zdola (nebo shora) nějakou konstantou K a že pro tuto mezní hodnotu bude splněno $\varrho(\boldsymbol{x}, \boldsymbol{y}) = K \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$ pro $\boldsymbol{x}, \boldsymbol{y} \in \left(\mathbb{R}_0^+\right)^N$ libovolné obrázky. Další požadovanou vlastností je také symetrie $\varrho(\boldsymbol{x}, \boldsymbol{y}) = \varrho(\boldsymbol{y}, \boldsymbol{x})$. Konkrétní volbu kritéria podobnosti je dále možné přizpůsobit konkrétní aplikaci a typu dat.

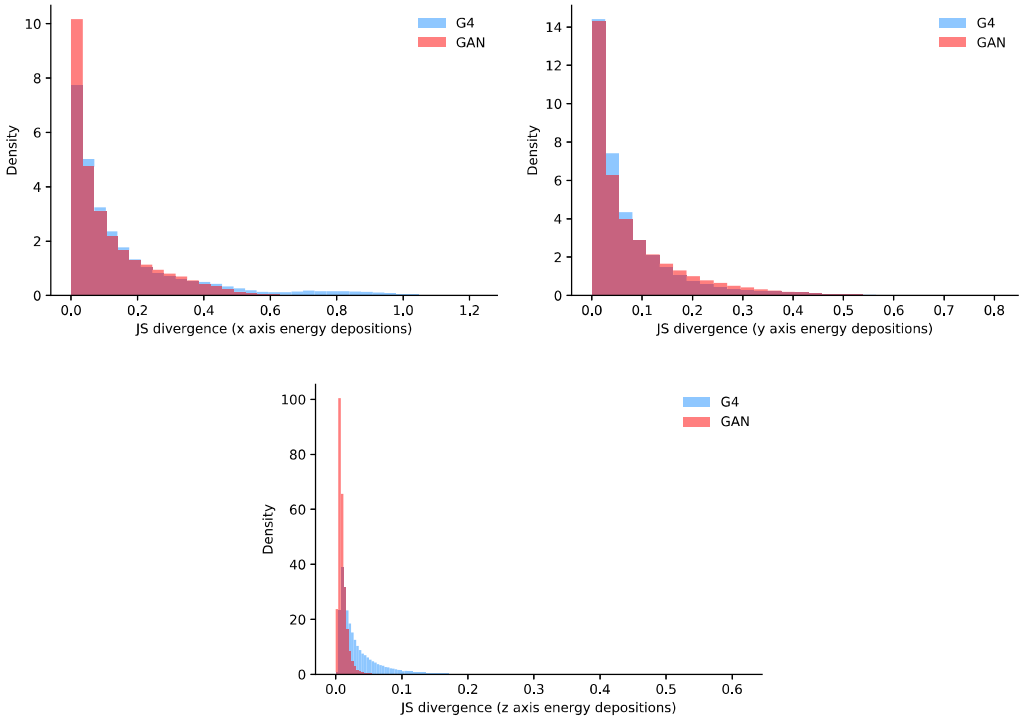
Stěžejní charakteristikou snímku elementární částice v kalorimetru je distribuce energie vyzářené částicí ve směrech hlavních os, v předchozí sekci nazvaná *shower shape*. Prvním kritériem, které pro odhad supportu modelu 3DGAN použijeme, bude tedy podobnost dvou snímků z hlediska distribuce energie částice ve směru os x, y a z. Podobnost těchto distribucí budeme měřit pomocí nenormalizované symetrické Jensenovy-Shannonovy divergence (zkráceně JS divergence), která je pro dvě rozdělení pravděpodobnosti P, Q na $\mathbb{R}$ definována jako

$$d'_{JS}(P, Q) = \frac{1}{2}\left[d'_{KL}\left(P, \frac{P+Q}{2}\right) + d'_{KL}\left(Q, \frac{P+Q}{2}\right)\right], \tag{5.7}$$

kde $d'_{KL}\left(P, \frac{P+Q}{2}\right) = \int_{\mathbb{R}} \left(P(x) \log\left(\frac{P(x)}{Q(x)}\right) - P(x) + Q(x)\right) \mathrm{d}x$ označuje nenormalizovanou Kullbackovu-Leiblerovu divergenci. Důvodem použití nenormalizované varianty JS divergence je skutečnost, že distribuce vyzářené energie ve směrech hlavních os není hustotou pravděpodobnosti a proto není normovaná na jedničku.

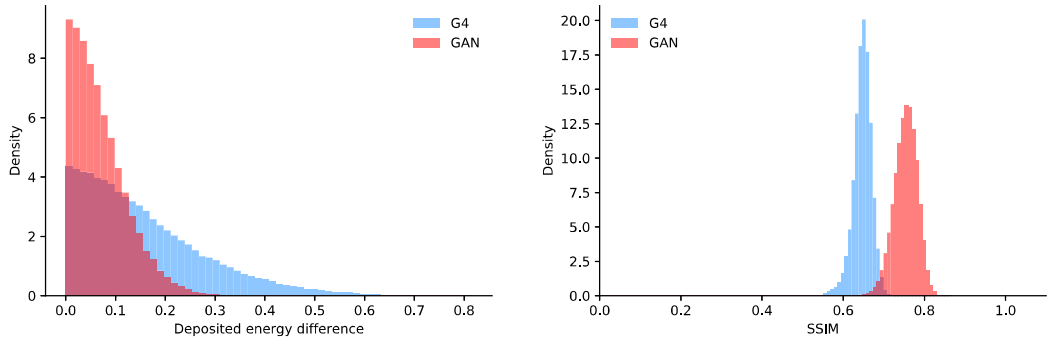
Rozdělení hodnot nenormalizované JS divergence distribucí energie podél hlavních os je znázorněno histogramy 5.1, které byly získány spočtením d'_{JS} pro všechny dvojice obrázků v množině 500 dat z 3DGAN a 500 MC simulací. U všech třech grafů lze pozorovat, že data modelu 3DGAN mají podobnější distribuce energie. U distribucí podél os x a y jsou rozdíly v histogramech pro 3DGAN data a MC data spíše drobné, pro distribuci ve směru osy z (směr letu částice) jsou odlišnosti v histogramech velmi zřetelné.

Jako druhé kritérium pro porovnání dvou snímků kalorimetru byla vybrána absolutní hodnota z rozdílu vyzářené energie, tedy ze součtu intenzit všech pixelů v obrázku. Označme tuto veličinu $E_{diff} = |E_{dep}(\boldsymbol{x}) - E_{dep}(\boldsymbol{y})|$, kde $E_{dep}(\boldsymbol{x}) = \sum_{i=1}^{N} x_i$ je součet intenzit všech pixelů.



Obrázek 5.1: Histogramy hodnot nenormalizované JS divergence distribucí energie podél hlavních os (po řadě x, y, z) pro 500 obrázků generovaných z 3DGAN (červená) nebo 500 MC simulací (modrá). Histogramy jsou normované na hustoty pravděpodobnosti.

Histogramy 5.2a ukazují přibližné rozdělení hodnot E_{diff} pro všechny dvojice z množiny 500 snímků z 3DGAN, respektive MC dat. Lze pozorovat, že data generovaná z 3DGAN mají podobnější hodnoty celkových vyzářených energií, přestože výstupy z 3DGAN byly podmíněny přesně stejnými hodnotami počáteční energie E_p jako simulace Monte Carlo.



(a) E_{diff} (b) SSIM

Obrázek 5.2: Histogramy hodnot (a) E_{diff} a (b) SSIM pro 500 snímků kalorimetru generovaných z 3DGAN (červená) nebo 500 MC simulací (modrá). Histogramy jsou normované na hustoty pravděpodobnosti.

Posledním kritériem, které bude aplikováno pro 3DGAN, je *index strukturní podobnosti* (angl. structural similarity index, zkráceně SSIM), který je často užívanou metrikou v úlohách zpracování obrazu [41]. SSIM pro dva obrázky $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ sestává ze třech komponent. První z nich je funkce $l(\boldsymbol{x}, \boldsymbol{y})$, která porovnává jas obrázků a je definovaná jako

$$l(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\mu_{\boldsymbol{x}}\mu_{\boldsymbol{y}} + C_1}{\mu_{\boldsymbol{x}}^2 + \mu_{\boldsymbol{y}}^2 + C_1}, \tag{5.8}$$

kde $\mu_{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} x_i$ je průměr intenzit pixelů v obrázku a C_1 je konstanta, která koriguje nestabilitu $l(\boldsymbol{x}, \boldsymbol{y})$ v případě, že $\mu_{\boldsymbol{x}}^2 + \mu_{\boldsymbol{y}}^2$ je blízké 0. Dále platí, že $C_1 = (K_1 L)^2$, kde L je rozsah hodnot intenzit pixelů (pro klasický šedotónový obrázek obvykle 255) a $K_1 \ll 1$ malá konstanta.

Snímky kalorimetru jsou ve srovnání s klasickými obrázky specifické značným rozsahem intenzit pixelů, které navíc nabývají nezvyklých hodnot, řádově $\sim 10^{-15} - 10^{-1}$. Stabilní hodnota L byla proto určena v [26] numerickými simulacemi, ze kterých vzešlo pro 3DGAN doporučení $L = 10^{-4}$.

Druhá komponenta SSIM porovnává kontrast obrázků $\boldsymbol{x}$, $\boldsymbol{y}$ a definujeme ji jako

$$c(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}} + C_2}{\sigma_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{y}}^2 + C_2}, \tag{5.9}$$

kde $\sigma_{\boldsymbol{x}} = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_{\boldsymbol{x}})^2\right)^{\frac{1}{2}}$ je nestranný odhad směrodatné odchylky intenzit pixelů v obrázku a $C_2 = (K_2 L)^2$ s konstantou $K_2 \ll 1$.

Poslední komponentou SSIM je funkce $s(\boldsymbol{x}, \boldsymbol{y})$, která má význam rozdílu v korelační struktuře obrázků $\boldsymbol{x}$, $\boldsymbol{y}$, a je definována jako

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sigma_{\boldsymbol{xy}} + C_3}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}} + C_3}, \tag{5.10}$$

kde $\sigma_{\boldsymbol{xy}} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_{\boldsymbol{x}})(y_i - \mu_{\boldsymbol{y}})$ a C_3 je malá konstanta.

Všechny tři výše definované komponenty SSIM nakonec spojíme do definice

$$\mathrm{SSIM}(\boldsymbol{x}, \boldsymbol{y}) = [l(\boldsymbol{x}, \boldsymbol{y})]^{\alpha} \cdot [c(\boldsymbol{x}, \boldsymbol{y})]^{\beta} \cdot [s(\boldsymbol{x}, \boldsymbol{y})]^{\gamma}, \tag{5.11}$$

ve které $\alpha > 0$, $\beta > 0$ a $\gamma > 0$. Takto definovaná funkce SSIM splňuje vlastnost symetrie $\mathrm{SSIM}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{SSIM}(\boldsymbol{y}, \boldsymbol{x})$, totožnosti $\mathrm{SSIM}(\boldsymbol{x}, \boldsymbol{y}) = 1 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$ a omezenosti shora $\mathrm{SSIM}(\boldsymbol{x}, \boldsymbol{y}) \leq 1$ pro libovolná $\boldsymbol{x}, \boldsymbol{y} \in \left(\mathbb{R}_0^+\right)^N$.

Pokud zvolíme $\alpha = \beta = \gamma = 1$ a $C_3 = C_2/2$, pak se SSIM zjednoduší na

$$\mathrm{SSIM}(\boldsymbol{x}, \boldsymbol{y}) = \frac{(2\mu_{\boldsymbol{x}}\mu_{\boldsymbol{y}} + C_1)(2\sigma_{\boldsymbol{xy}} + C_2)}{(\mu_{\boldsymbol{x}}^2 + \mu_{\boldsymbol{y}}^2 + C_1)(\sigma_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{y}}^2 + C_2)}. \tag{5.12}$$

V praxi není hodnota SSIM počítána pro celé obrázky, ale počítá se lokálně pro menší výřez obrázku, který se posouvá po jednotlivých pixelech. Výsledný SSIM je získán jako průměr lokálně spočtených hodnot. Pro třírozměrný snímek kalorimetru se k rozměru ve směru osy z (neboli ve směru pohybu částice) chováme jako k barevnému kanálu u RGB obrázků. SSIM je tedy spočítán pro každý průřez v rovině xy zvlášť a výsledné hodnoty jsou opět zprůměrovány.

Histogram 5.2b zobrazuje přibližné rozdělení hodnot SSIM spočtených pro 500 snímků z 3DGAN a 500 MC simulací. Rozdělení hodnot SSIM pro tyto dva soubory dat se téměř nepřekrývá, jasně lze také vidět, že data generovaná z 3DGAN jsou vzhledem k SSIM vzájemně výrazně podobnější než MC simulace.

Pokud již máme zvolené funkce pro měření podobnosti objektů z GAN, je potřeba pro tyto funkce určit hraniční hodnoty určení duplikátů. Pokud bude hodnota funkce podobnosti pro dvojici obrázků pod touto hranicí, příslušné objekty označíme za duplikáty. Pro případ snímků kalorimetru byla ke stanovení této hranice využita trénovací MC data. Nejprve bylo vybráno 500 obrázků z trénovací množiny. Pro všechny možné dvojice snímků v rámci této množiny byly spočítány hodnoty výše definovaných kritérií podobnosti, celkem byly výpočty provedeny pro 124 750 různých dvojic. Následně byl pro každé z kritérií podobnosti spočten výběrový α-kvantil ze získaných hodnot, a to pro $\alpha \in \{0.05, 0.02, 0.01\}$ v případě porovnávání distribucí energií podél os a rozdílů vyzářené energie, pro $\alpha \in \{0.95, 0.98, 0.99\}$ v případě SSIM. Na rozdíl od ostatních kritérií vyšší hodnota SSIM znamená větší podobnost obrázků. Hodnoty nalezených kvantilů jsou uvedeny v tabulce 5.1.

Tabulka 5.1: Tabulka výběrových α-kvantilů různých kritérií podobnosti pro MC data.

α	0,05	0,02	0,01
JS div. (osa x)	0,0109	0,0083	0,0072
JS div. (osa y)	0,0076	0,0063	0,0056
JS div. (osa z)	0,0065	0,0053	0,0046
E_{diff}	0,0115	0,0047	0,0024
SSIM	0,6806	0,6882	0,6934

Následně bylo generováno 500 snímků z 3DGAN a opět byla pro všechny dvojice vyčíslena kritéria podobnosti. Pro tuto množinu dat pak bylo u každého kritéria vyhodnoceno,

kolik duplikátních dvojic je v ní možné při hranici určené příslušným výběrovým kvantilem z MC pozorování nalézt. Tyto počty jsou k nahlédnutí v tabulce 5.2. Pro $\alpha = 0,05$ jsou počty nalezených duplikátů v množině 500 snímků z GAN poměrně vysoké. Při takto stanovené hranici pro definici duplikátů se tedy dá předpokládat vysoká pravděpodobnost nalezení duplikátních obrázků i pro velmi malé množiny dat z 3DGAN. Naopak při stanovení hranice pro určení duplikátu výběrovým kvantilem pro $\alpha = 0,01$ by pravděpodobnost nalezení duplikátní dvojice pro MC data byla i pro velké množiny obrázků velmi malá. V simulacích pro odhadnutí supportu uvedených v další sekci byl proto jako hraniční hodnota definice duplikátních snímků použit výběrový 0,02-kvantil MC dat.

Tabulka 5.2: Počty nalezených duplikátů pro množinu o 500 snímcích z 3DGAN z hlediska různých kritérií.

α	0,05	0,02	0,01
JS div. (osa x)	15 221	9 658	7 029
JS div. (osa y)	14 518	10 189	7 967
JS div. (osa z)	44 835	29 044	20 377
E_{diff}	13 408	5 482	2 877
SSIM	123 495	122 268	118 517

Z této úvahy je již jasné, že počet nalezených duplikátů, a tedy i odhad velikosti supportu, bude přímo záviset na stanovení hraniční hodnoty pro definici duplikátů vzhledem k danému kritériu podobnosti ϱ. Výsledek tohoto postupu pro 3DGAN tak bude sloužit spíše pro porovnání výsledků generativní sítě s trénovacími MC daty.

Při použití principu narozeninového problému pro odhad supportu GAN v [10] byla tato metoda aplikována na generativní modely napodobující fotografie tváří (dataset CelebA [45]) nebo jiných objektů (CIFAR-10 [27]). U této úlohy autoři nejprve určili podobnostní kritéria, ta vyčíslili pro zkoumanou množinu obrázků, nalezli několik nejpodobnějších dvojic a ty pak byly přezkoumány lidským hodnotitelem, který rozhodl o jejich shodnosti nebo rozdílnosti.

Stejný postup nelze pro 3DGAN aplikovat, protože snímky kalorimetru jsou trojrozměrné, takže lidský hodnotitel by musel o shodnosti obrázků rozhodovat na základě dvourozměrných projekcí, kterými se ale část informace o rozložení vyzářené energie ztratí. Navíc vzhledem k objemu dat, která s v HEP zpracovávají, je žádoucí mít k dispozici metodu, která funguje pokud možno automaticky.

5.3 Výsledky odhadu supportu modelu 3DGAN

MC simulace kalorimetru a stejně tak syntetické snímky z 3DGAN jsou generovány pro elektrony s různou počáteční energií E_p a různým úhlem vstupu θ do kalorimetru. Odezva kalorimetru na částici se pro různé energie a úhly značně liší. Všechny simulační experimenty prezentované v této kapitole byly proto prováděny pouze pro simulace elektronů s počáteční energií $E_p = 100 \pm 2,5\text{GeV}$ a úhlem $\theta = 90° \pm 2,5°$.

Následující kroky popisují simulaci, jejímž cílem bylo získat odhad velikosti supportu modelu 3DGAN.

1. Z modelu 3DGAN generujeme s snímků kalorimetru $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(s)}$.

2. Spočteme hodnoty vybraných kritérií podobnosti pro všechny možné dvojice obrázků získané v kroku 1.

3. Vyhodnotíme, jestli byla v dané množině nalezena alespoň jedna dvojice duplikátních obrázků.

4. Kroky 1.-3. provedeme v $k = 1\,000$ opakováních.

5. Spočteme odhad pravděpodobnosti nalezení duplikátů v množině dat o velikosti s jako podíl počtu iterací, ve kterých byly nalezeny duplikáty, a celkového počtu iterací k.

6. Zvětšujeme počet snímků s zkoumané množiny a opakujeme kroky 1.-5., dokud není odhad pravděpodobnosti nalezení duplikátních snímků alespoň 50 %.

Nejmenší s^*, pro které při tomto postupu bude odhad pravděpodobnosti nalezení duplikátů alespoň 50 %, pak poskytuje odhad velikosti supportu modelu 3DGAN, který je určen jako s^{*2}.

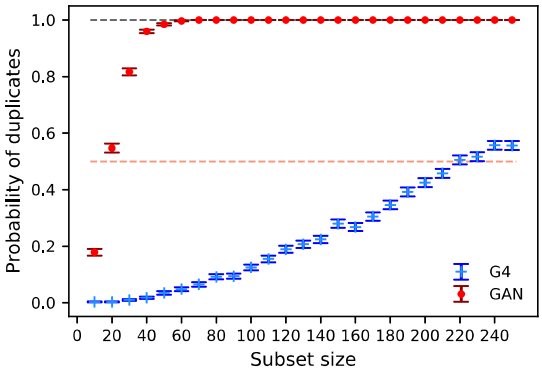
Stejným způsobem lze provést také odhad velikosti supportu trénovacích MC dat, jen v prvním kroku vybíráme množinu s pozorování z trénovacích dat.

Graf 5.3a zobrazuje odhady pravděpodobností nalezení duplikátů pro množiny o různých počtech pozorování s (označeno *subset size*). Červené body odpovídají výsledkům simulací pro 3DGAN, modré body MC datům. Červenou přerušovanou čarou je označena hranice 50 %. Při této simulaci byly jako kritéria podobnosti obrázků použity pouze rozdíly v distribucích energie podél hlavních os vyjádřené nenormalizovanou JS divergencí. Hladina 50% pravděpodobnosti nalezení duplikátů byla v případě 3DGAN dat překročena už pro množinu o 20 obrázcích, podle principu narozeninového problému je tedy velikost supportu naučeného rozdělení pouze 400. Pro MC data bylo pravděpodobnosti 50 % dosaženo až pro dataset o 220 pozorováních, odhad velikosti supportu je tedy 48 400.

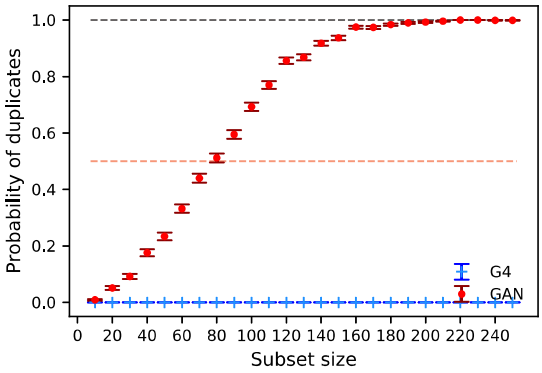
V dalším experimentu bylo ke kritériu shodnosti distribuce energie podél os přidáno také kritérium shodnosti celkové vyzářené energie, které je vyjádřeno funkcí E_{diff}. Výsledky simulací se přidáním tohoto kritéria výrazně změnily, jak lze nahlédnout v grafu 5.3b. Při simulacích pro 3DGAN data byla hranice 50% pravděpodobnosti nalezení duplikátů překročena pro 80 snímků, z této simulace tedy plyne odhad velikosti supportu 6 400 pro 3DGAN. Výsledky simulací pro MC data se k hranici 50 % ani nepřiblížily, odhad velikosti supportu tedy nelze získat.

Při poslední simulaci byl k předchozím kritériím podobnosti přidán také SSIM, výsledky jsou zobrazeny v grafu 5.3c. Dodáním tohoto kritéria se odhady pravděpodobností nalezení duplikátů prakticky nezměnily, jedná se ale o očekávaný výsledek. V grafu 5.2b lze nahlédnout, že hodnoty SSIM pro dvojice dat z 3DGAN jsou výrazně vyšší než pro dvojice MC simulací, téměř všechny dvojice 3DGAN obrázků tedy splní podmínku pro označení za duplikátní vzhledem ke kritériu SSIM.

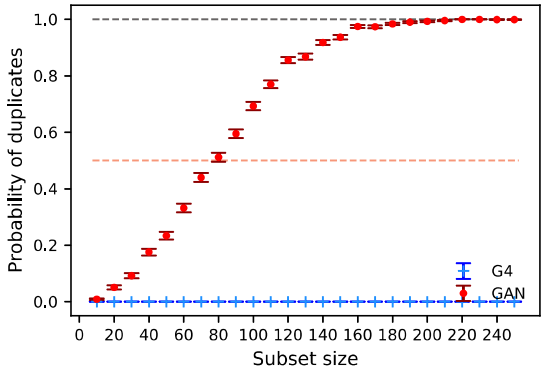
Na obrázku 5.4 jsou k nahlédnutí 2D projekce dvou snímků kalorimetru generovaných z 3DGAN, které byly při použití všech kritérií podobnosti současně určeny jako duplikátní. Přestože se podle výsledků simulací pro MC data mohlo aplikování všech kritérií současně zdát poměrně striktní, mezi oběma zobrazenými simulacemi jsou i pouhým okem viditelné rozdíly v jejich 2D projekcích a za duplikáty bychom je nejspíš nepovažovali.



(a) JS divergence (x, y, z)

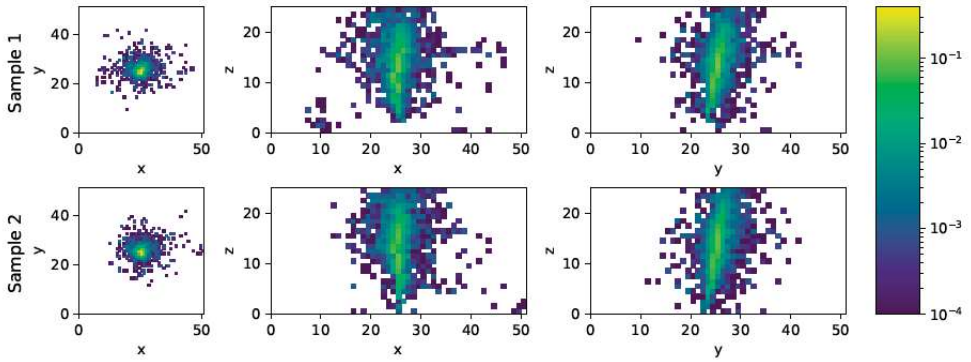


(b) JS div. a E_{diff}



(c) JS div., E_{diff} a SSIM

Obrázek 5.3: Odhady pravděpodobnosti výskytu duplikátu v množině dat o velikosti s (*subset size*). Červená křivka odpovídá datům z 3DGAN, modrá křivka MC simulacím. Použitá kritéria podobnosti jsou uvedena v popiscích grafů.



Obrázek 5.4: 2D projekce snímků kalorimetru. Intenzita barvy jednotlivých bodů odpovídá energii vyzářené částicí v daném místě v jednotkách GeV. Oba snímky byly generovány z 3DGAN a byly vyhodnoceny jako duplikáty při použití všech popsaných kritérií podobnosti současně.

Získané výsledky potvrdily, že uvedený automatizovaný přístup k definici duplikátních dat není vhodný přímo pro získání odhadu velikosti supportu naučeného rozdělení modelu 3DGAN ani skutečného rozdělení trénovacích dat. Výsledek je pro data ze spojitého rozdělení přímo závislý na způsobu, jakým se duplikátní pozorování definují. Na výsledný odhad supportu má vliv jak volba kritérií podobnosti dat, tak stanovení hraničních hodnot.

Metoda se však ukázala jako užitečná pro porovnání generativního modelu a trénovacích dat, případně dvou různých generativních modelů. Jasně prokázala, že data generovaná modelem 3DGAN jsou vzájemně mnohem podobnější než MC data a při opakovaném generování z 3DGAN modelu se obrázky začnou brzy opakovat. Tento závěr rozporuje výsledky uvedené v [16], které naznačovaly, že 3DGAN velmi dobře napodobuje rozdělení MC dat. Otevřel se tím další prostor k vylepšení generativního modelu 3DGAN. Díky obecnosti celého přístupu je navíc možné volbou kritérií tuto metodu uzpůsobit přesně podle požadavků konkrétní úlohy a typu dat.

Závěr

V této práci bylo nahlédnuto na problematiku simulací v HEP ze dvou odlišných úhlů. V kapitole 1 byla nejprve vysvětlena důležitost simulací při bádání na poli částicové fyziky a byly shrnuty základní informace o Monte Carlo simulacích. Zároveň byly poskytnuty informace o vývoji GANs jako alternativní metodě pro rychlé generování simulací kalorimetrů.

Následující kapitola 2 nastínila problém testování homogenity vážených souborů dat, který se v HEP analýzách pravidelně objevuje. Byl poskytnut teoretický základ k neparametrickým testům homogenity a jádrovým odhadům hustot. Dále byla provedena rešerše stávajících přístupů k testu homogenity vážených dat se zaměřením na neparametrické testy s testovací statistikou modifikovanou pro vážená data. Druhou zkoumanou metodou byly testy využívající re-arranging, který umožňuje transformovat množinu vážených pozorování na nevážená.

V kapitole 3 byla navržena vlastní modifikace pro ověření homogenity vážených dat. Tato metoda využívá WKDE nebo WAKDE k získání odhadu hustoty pravděpodobnosti vážených dat, ze které jsou následně generována nevážená pozorování. Poté byly provedeny numerické simulace testů homogenity vážených dat s cílem ověřit použitelnost čtyř různých zobecnění Kolmogorovova-Smirnovova testu pro vážená data: testu s WKDE, testu s WAKDE, testu s modifikovanou statistikou a testu s metodou re-arranging. Zkoumána byla především pravděpodobnost chyby I. druhu, respektive dodržení hladiny významnosti testu α za platnosti H_0. Ve vybraných simulacích byla porovnávána také síla testu pro různou úroveň znečištění dat jednoho ze souborů.

Pro test s modifikovanou statistikou došlo ve srovnání s [37] k výraznému rozšíření ověření jeho funkčnosti. Podařilo se ukázat, že tato varianta KS testu funguje poměrně spolehlivě nejen pro data z normálního rozdělení, ale také pro soubory dat z logistického, lognormálního, gamma a Weibullova rozdělení a s váhami z rozdělení Beta nebo uniformního. Numerickými simulacemi bylo navíc potvrzeno, že tento test zachovává asymptotickou hladinu významnosti nejen při testování váženého souboru dat s neváženým, ale také pro dva vážené soubory.

Použití metody re-arranging pro testování homogenity dvou vážených souborů je také novým výsledkem. Pomocí simulací se podařilo ukázat, že při použití dvou vážených souborů klesla P(chyby I. druhu) tohoto testu téměř k nule. To zjevně ovlivnilo také sílu testu, která byla výrazně nižší, než pro ostatní srovnávané varianty testů. Metodu re-arrangingu tedy nelze na základě výsledků této práce doporučit pro spolehlivé testování homogenity vážených dat. Její použití v [14] jako referenční metody k ověření fungování testů s modifikovanou statistikou nebylo podle zjištěných výsledků dobrou volbou.

Pro testy s využitím WKDE a WAKDE bylo numerickými simulacemi zjištěno, že odhad P(chyby I. druhu) více než trojnásobně překročil zvolenou hladinu významnosti

testu $\alpha = 0,05$. Bylo ověřeno, že toto chování není způsobeno mechanismem generování dat z WKDE nebo WAKDE. Příčinou je tedy zřejmě samotná chyba jádrového odhadu. Dále byla provedena srovnávací simulace na datech z rozdělení $\mathrm{N}(0,1)$. KDE test byl porovnán s nově navrženým MLE testem, u kterého byla předpokládána znalost normální rodiny rozdělení a z dat byly odhadovány pouze parametry normálního rozdělení metodou MLE. Získané výsledky ukázaly, že ani znalost parametrické rodiny rozdělení dat nepomůže výrazně snížit odhad P(chyby I. druhu). Při odhadování rozdělení z dat a následném generování nových pozorování z tohoto odhadu tedy dochází k výraznému kumulování chyby, kterou KS test dokáže zachytit a nelze již využít asymptotické rozdělení testovací statistiky klasického KS testu ke stanovení kritické hodnoty pro zamítnutí H_0.

Na závěr bylo na základě analýzy distribuce p-hodnot testu s jádrovými odhady navrhnuto řešení umožňující test s jádrovým odhadem přibližně naladit na požadovanou P(chyby I. druhu) $= \alpha$. Fungování navrženého řešení bylo demonstrováno pro vybrané rozdělení vážených pozorování.

Menší část práce pak byla věnována sítím GAN a jejich využití pro simulace v HEP. V kapitole 4 byly nejprve shrnuty základní informace o vícevrstvých neuronových sítích a principu fungování GAN. Následně byl představen model 3DGAN, který slouží ke generování snímků interakce elektronu v EM kalorimetru. Pro tento konkrétní model byla v kapitole 5 implementována metoda evaluace kvality generovaných dat založená na principu problému narozenin. V rozporu s předchozími výsledky analýzy kvality sítě 3DGAN [16] bylo ukázáno, že v porovnání s MC simulacemi, na kterých byl tento model trénován, mají simulace z 3DGAN výrazně menší variabilitu. Byl tak otevřen další prostor pro návrhy na vylepšení tohoto modelu. Výsledky z této části práce byly prezentovány na *4th Inter-experiment Machine Learning Workshop* v CERN v říjnu 2020 a dále na workshopu ML4PS (*Machine Learning and Physical Sciences*) organizovaného v rámci konference NeurIPS v prosinci 2020.

Literatura

[1] ABAZOV, V. M. et al., Measurement of the inclusive $t\bar{t}$ production cross section in $p\overline{p}$ collisions at $\sqrt{s} = 1 : 96$TeV and determination of the top quark pole mass, *Phys. Rev. D*, vol. 94, 2016.

[2] ABRAMSON, Ian S. On Bandwidth Variation in Kernel Estimates-A Square Root Law. *The Annals of Statistics* [online]. 1982, 10(4) [cit. 2021-4-25]. ISSN 0090-5364. Dostupné z: doi:10.1214/aos/1176345986

[3] AGGARWAL, Charu C. *Neural networks and deep learning: a textbook*. Cham: Springer, [2018].

[4] ALBERTSSON, Kim, Piero ALTOE, Dustin ANDERSON, et al. Machine Learning in High Energy Physics Community White Paper. *Journal of Physics: Conference Series* [online]. 2018, **1085** [cit. 2021-5-3]. ISSN 1742-6588. Dostupné z: doi:10.1088/1742-6596/1085/2/022008

[5] ALBRECHT, Johannes, Antonio Augusto ALVES, Guilherme AMADIO, et al. A Roadmap for HEP Software and Computing R&D for the 2020s. *Computing and Software for Big Science* [online]. 2019, **3**(1) [cit. 2021-4-3]. ISSN 2510-2036. Dostupné z: doi:10.1007/s41781-018-0018-8

[6] ALLISON, J., K. AMAKO, J. APOSTOLAKIS, et al. Recent developments in Geant4. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* [online]. 2016, 835, 186-225 [cit. 2021-04-11]. ISSN 01689002. Dostupné z: doi:10.1016/j.nima.2016.06.125

[7] AMADIO, G., A. ANANYA, J. APOSTOLAKIS, et al. GeantV. *Computing and Software for Big Science* [online]. 2021, 5(1) [cit. 2021-04-11]. ISSN 2510-2036. Dostupné z: doi:10.1007/s41781-020-00048-6

[8] ARJOVSKY, M., S. CHINTALA a L. BOTTOU.Wasserstein Generative Adversarial Networks. *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:214-223, 2017.

[9] ARORA, S., R. GE, Y. LIANG. Generalization and Equilibrium in Generative Adversarial Nets (GANs). *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:224-232, 2017.

[10] ARORA, Sanjeev a Yi ZHANG. *Do GANs actually learn the distribution? An empirical study*, 2017. arXiv:1706.08224 [cs.LG].

[11] AURISANO, A., A. RADOVIC, D. ROCCO, et al. A convolutional neural network neutrino event classifier. *Journal of Instrumentation* [online]. 2016, **11**(09), P09001-P09001 [cit. 2021-4-14]. ISSN 1748-0221. Dostupné z: doi:10.1088/1748-0221/11/09/P09001

[12] BELLM, Johannes, Gavin BEWICK, Silvia FERRARIO RAVASIO, et al. Herwig 7.2 release note. *The European Physical Journal C* [online]. 2020, **80**(5) [cit. 2021-3-26]. ISSN 1434-6044. Dostupné z: doi:10.1140/epjc/s10052-020-8011-x

[13] BOUŘ, Petr. *Vývoj statistických neparametrických a divergenčních metod pro zpracování dat z experimentů DØ a NOvA*. Praha, 2016. Diplomová práce. České vysoké učení technické v Praze. Fakulta jaderná a fyzikálně inženýrská.

[14] BOUŘ, Petr a Václav KŮS. Computer simulation on homogeneity testing for weighted data sets used in HEP. *Journal of Physics: Conference Series* [online]. 2018, **1085** [cit. 2021-3-2]. ISSN 1742-6588. Dostupné z: doi:10.1088/1742-6596/1085/4/042002

[15] BRINK, David. A (probably) exact solution to the Birthday Problem. *The Ramanujan Journal* [online]. 2012, **28**(2), 223-238 [cit. 2021-5-3]. ISSN 1382-4090. Dostupné z: doi:10.1007/s11139-011-9343-9

[16] BRITO DA ROCHA, Ricardo, Federico CARMINATI, Gulrukh KHATTAK, et al. Fast simulation of electromagnetic particle showers in high granularity calorimeters. *EPJ Web of Conferences* [online]. 2020, 245 [cit. 2021-04-13]. ISSN 2100-014X. Dostupné z: doi:10.1051/epjconf/202024502034

[17] CERN. High Energy Physics simulations. *https://lhcathome.web.cern.ch/* [online]. © 2021 [cit. 2021-03-26]. Dostupné z https://lhcathome.web.cern.ch/projects/test4theory/high-energy-physics-simulations

[18] CERN. High-Luminosity LHC. *home.cern* [online]. © 2021 [cit. 2021-03-26]. Dostupné z https://home.cern/science/accelerators/high-luminosity-lhc

[19] CERN. The Standard Model. *home.cern* [online]. © 2021 [cit. 2021-03-26]. Dostupné z https://home.cern/science/physics/standard-model

[20] DE OLIVEIRA, Luke, Michela PAGANINI a Benjamin NACHMAN. Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis. *Computing and Software for Big Science* [online]. 2017, **1**(1) [cit. 2021-4-16]. ISSN 2510-2036. Dostupné z: doi:10.1007/s41781-017-0004-6

[21] DEGROOT, Morris H. a Mark J. SCHERVISH: *Probability and Statistics 4th Edition.* Pearson, 2012. ISBN-13: 978-0321500465.

[22] DEVROYE, Luc a László GYÖRFI. *Nonparametric Density Estimation: the L_1 view.* New York: John Wiley & Sons, 1985.

[23] FREEMAN, Peter, Stephen DOE, Aneta SIEMIGINOWSKA, Jean-Luc STARCK a Fionn D. MURTAGH. [online]. In: . 2001-11-1, s. 76-87 [cit. 2021-3-26]. Dostupné z: doi:10.1117/12.447161

[24] GOODFELLOW, Ian, Y. BENGIO a A.COURVILLE. Deep Learning. MIT Press, 2016.

[25] GOODFELLOW, Ian, Jean POUGET-ABADIE, Mehdi MIRZA, Bing XU, David WARDE-FARLEY, Sherjil OZAIR, Aaron COURVILLE a Yoshua BENGIO. Generative adversarial networks. *Communications of the ACM* [online]. 2020, **63**(11), 139-144 [cit. 2021-4-16]. ISSN 0001-0782. Dostupné z: doi:10.1145/3422622

[26] KHATTAK, Gul Rukh, Sofia VALLECORSA, Federico CARMINATI a Gul Muhammad KHAN. Particle Detector Simulation using Generative Adversarial Networks with Domain Related Constraints. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* [online]. IEEE, 2019, 2019, s. 28-33 [cit. 2021-5-3]. ISBN 978-1-7281-4550-1. Dostupné z: doi:10.1109/ICMLA.2019.00014

[27] KRIZHEVSKY, Alex. Learning multiple layers of features from tiny images [online]. 2009 [cit. 2021-04-12]. DOstupné z https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf

[28] KUBŮ, Miroslav. *Metody strojového učení v částicové fyzice*. Diplomová práce, České Vysoké Učení Technické v Praze, Praha, 2021.

[29] MARSAGLIA, George, Wai Wan TSANG a Jingbo WANG. Evaluating Kolmogorov's Distribution. *Journal of Statistical Software* [online]. 2003, **8**(18) [cit. 2021-5-3]. ISSN 1548-7660. Dostupné z: doi:10.18637/jss.v008.i18

[30] MASSEY, Frank J. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* [online]. 1951, **46**(253) [cit. 2021-5-3]. ISSN 01621459. Dostupné z: doi:10.2307/2280095

[31] MEDIUM. Introduction to Generative Adversarial Networks (GANs) *medium.com* [online]. [cit. 2021-4-16]. Dostupné z https://medium.com/@kraken2404/introduction-to-generative-adversarial-networks-gans-89095151cd3a

[32] PAGANINI, Michela, Luke DE OLIVEIRA a Benjamin NACHMAN. Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters. *Physical Review Letters* [online]. 2018, **120**(4) [cit. 2021-4-16]. ISSN 0031-9007. Dostupné z: doi:10.1103/PhysRevLett.120.042003

[33] RADFORD Alec, Luke METZ a soumith CHINTALA. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, 2015. arXiv:1511.06434 [cs.LG].

[34] SILVERMAN, B. W. *Density estimation for statistics and data analysis*. Boca Raton: Chapman & Hall/CRC, 1998. ISBN 0412246201.

[35] SJÖSTRAND, Torbjörn, Stephen MRENNA a Peter SKANDS. PYTHIA 6.4 physics and manual. *Journal of High Energy Physics* [online]. 2006, 2006(05), 026-026 [cit. 2021-3-26]. ISSN 1029-8479. Dostupné z: doi:10.1088/1126-6708/2006/05/026

[36] TRUSINA, Jakub. *Application of statistical hypothesis testing to datasets from the ATLAS experiment.* Bakalářská práce, České Vysoké Učení Technické v Praze, Praha, 2017.

[37] TRUSINA, Jakub. *Aplikace testů homogenity a klasifikace částic v neutrinové fyzice.* Praha, 2019. Diplomová práce. České vysoké učení technické v Praze. Fakulta jaderná a fyzikálně inženýrská.

[38] TRUSINA, Jakub, Jiří FRANC a Adam NOVOTNÝ. Generalization of Homogeneity Tests for Weighted Samples and Their Implementation in ROOT. *Journal of Physics: Conference Series* [online]. 2020, **1525** [cit. 2021-3-2]. ISSN 1742-6588. Dostupné z: doi:10.1088/1742-6596/1525/1/012109

[39] VAN DER VAART, A. Asymptotic Statistics. *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, 1998. ISBN 0-521-78450-6.

[40] VALLECORSA, S. Generative models for fast simulation. *Journal of Physics: Conference Series* [online]. 2018, 1085 [cit. 2021-4-16]. ISSN 1742-6588. Dostupné z: doi:10.1088/1742-6596/1085/2/022005

[41] WANG, Z., A.C. BOVIK, H.R. SHEIKH a E.P. SIMONCELLI. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* [online]. 2004, **13**(4), 600-612 [cit. 2021-5-3]. ISSN 1057-7149. Dostupné z: doi:10.1109/TIP.2003.819861

[42] WIKIPEDIA. Standard Model of Elementary Particles. *https://en.wikipedia.org/* [online]. [cit. 2021-03-26]. Dostupné z https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg

[43] ZABOROWSKA, A. Geant4 fast and full simulation for Future Circular Collider studies. *Journal of Physics: Conference Series* [online]. 2017, **898** [cit. 2021-4-16]. ISSN 1742-6588. Dostupné z: doi:10.1088/1742-6596/898/4/042053

[44] ZECH, Gunter. Comparing statistical data to monte carlo simulationparameter fitting and unfolding, *INIS*, vol. 27, pp. 95–113, June 1995.

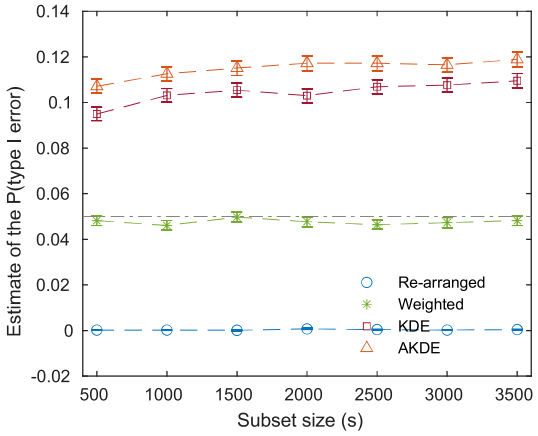
[45] ZIWEI Liu, P. LUO, X. WANG, a X. TANG. *Deep learning face attributes in the wild*, 2015. arXiv:1411.7766 [cs.CV].

Příloha A

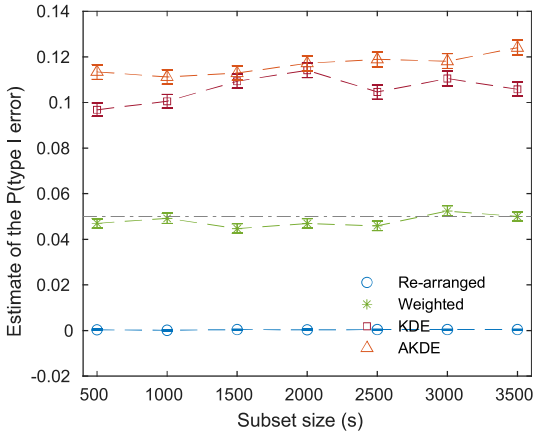
Testy homogenity

A.1 Testy dvou vážených souborů dat

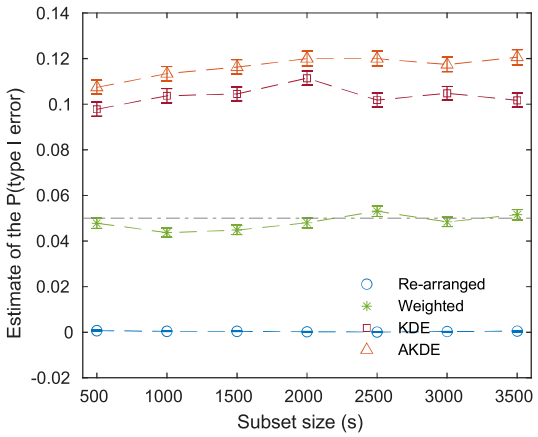
A.1.1 Rozdělení vah Beta(2,4)



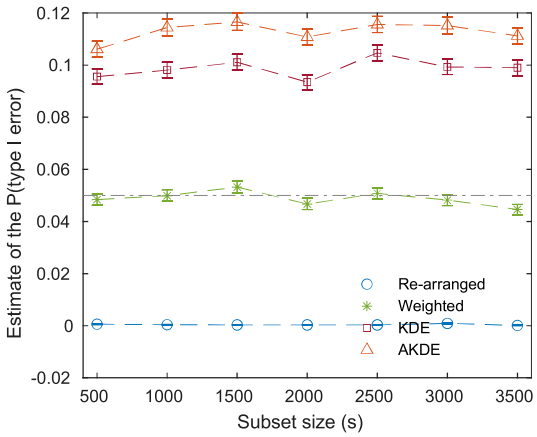
(a) Normální rozdělení



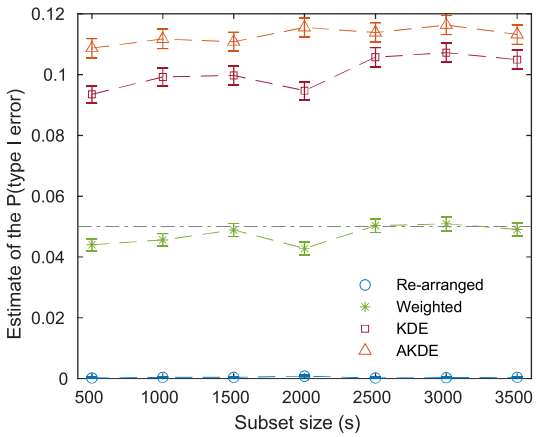
(b) Logistické rozdělení



(c) Lognormalní rozdělení



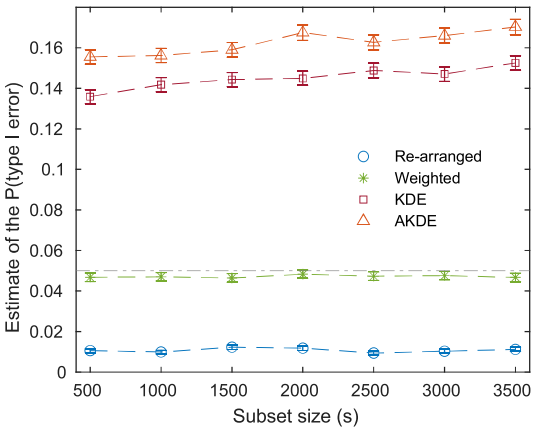
(d) Gamma rozdělení



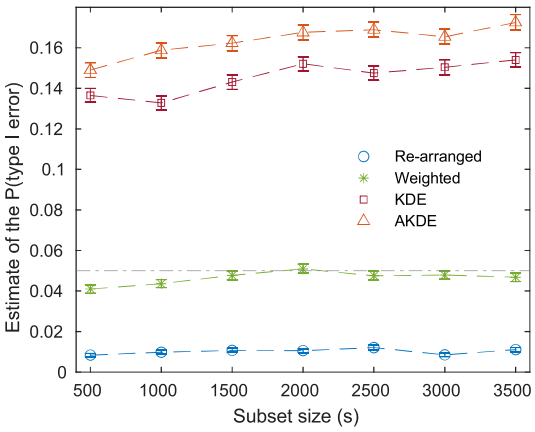
(e) Weibullovo rozdělení

Obrázek A.1: Odhad pravděpodobnosti chyby I. druhu pro testy dvou vážených souborů ze stejného rozdělení. Pozorování X, resp. Y byla generována z rozdělení uvedeného v popisu grafu. Váhy byly generovány z rozdělení Beta$(2, 4)$.

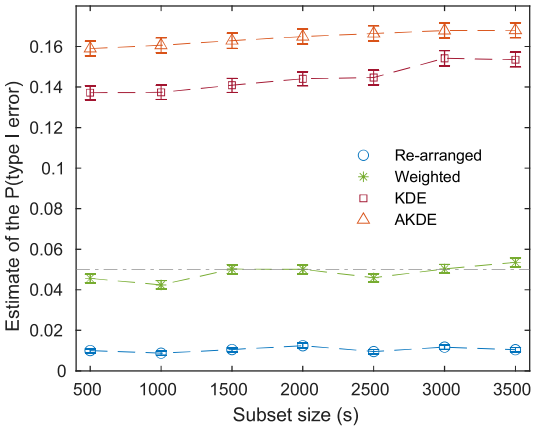
A.1.2 Rozdělení vah Beta(0.7,0.7)



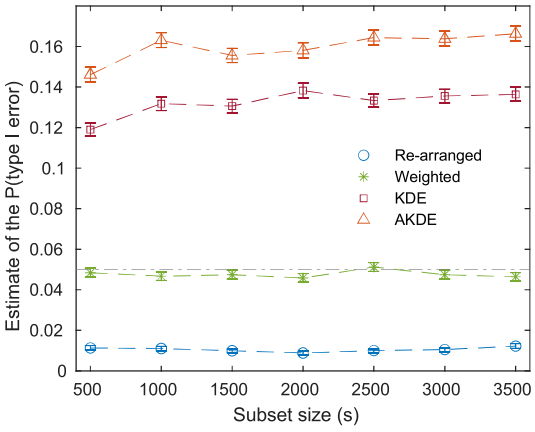
(a) Normální rozdělení



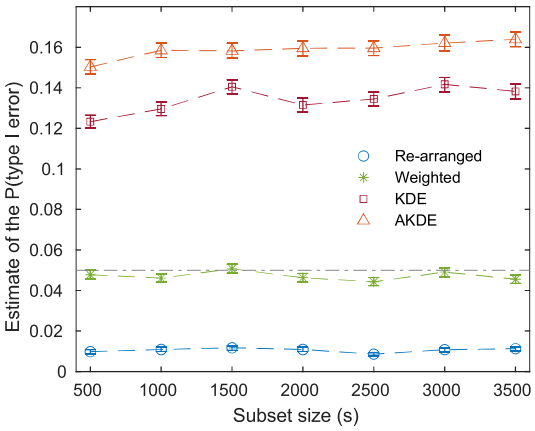
(b) Logistické rozdělení



(c) Lognormalní rozdělení



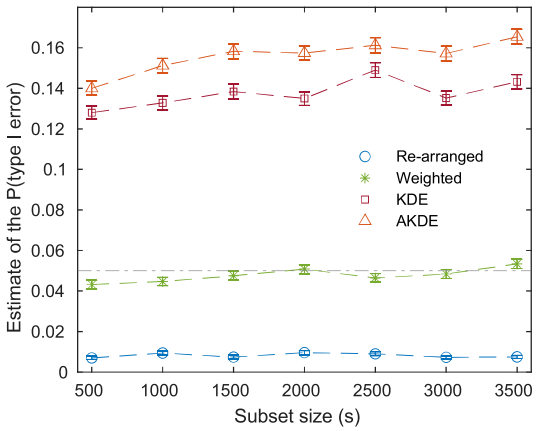
(d) Gamma rozdělení



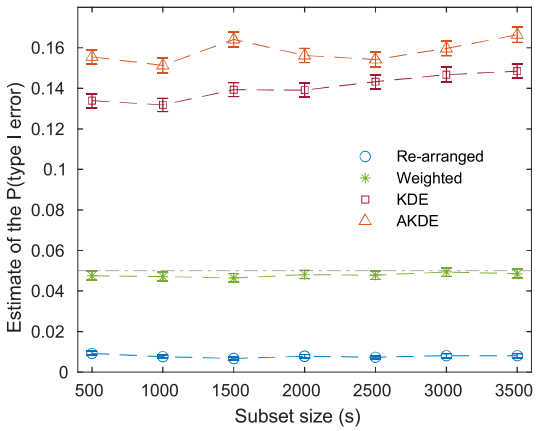
(e) Weibullovo rozdělení

Obrázek A.2: Odhad pravděpodobnosti chyby I. druhu pro testy dvou vážených souborů ze stejného rozdělení. Pozorování X, resp. Y byla generována z rozdělení uvedeného v popisu grafu. Váhy byly generovány z rozdělení Beta$(0.7, 0.7)$.

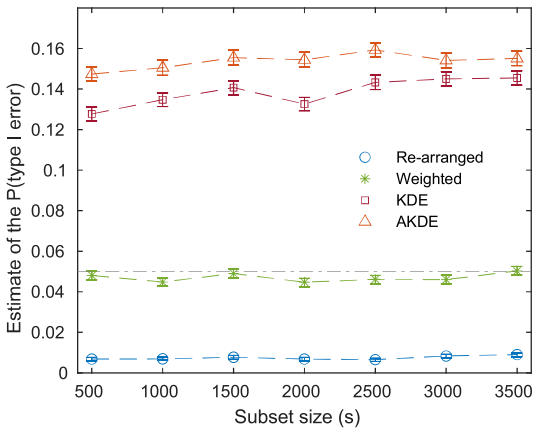
A.1.3 Rozdělení vah U(0,1)



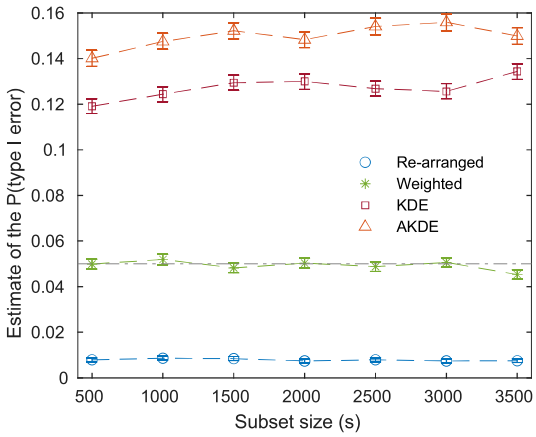
(a) Normální rozdělení



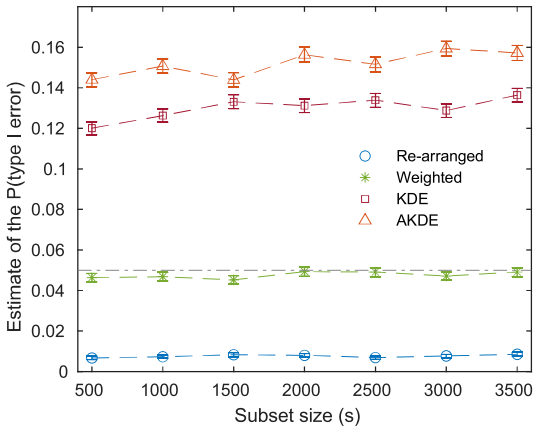
(b) Logistické rozdělení



(c) Lognormalní rozdělení



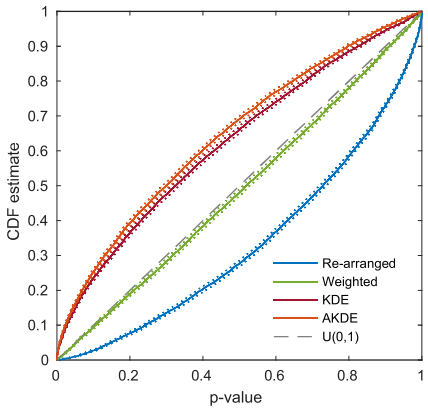
(d) Gamma rozdělení



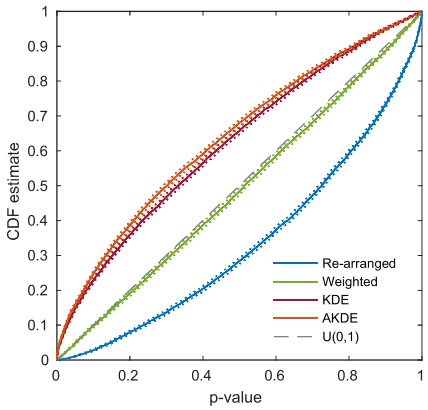
(e) Weibullovo rozdělení

Obrázek A.3: Odhad pravděpodobnosti chyby I. druhu pro testy dvou vážených souborů ze stejného rozdělení. Pozorování X, resp. Y byla generována z rozdělení uvedeného v popisu grafu. Váhy byly generovány z rozdělení $\mathrm{U}(0,1)$.

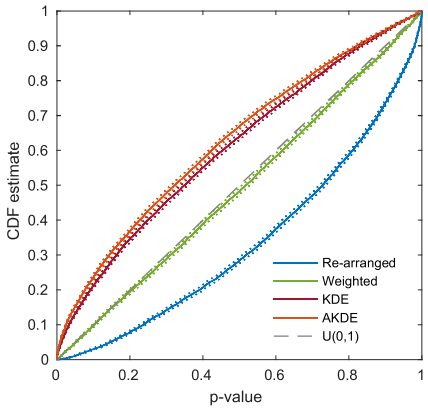
A.2 Empirické distribuční funkce p-hodnot



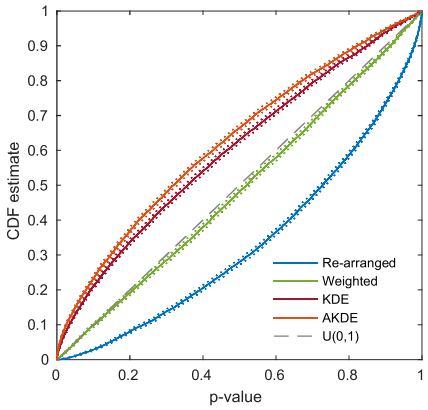
(a) Logistické rozdělení



(b) Lognormalní rozdělení



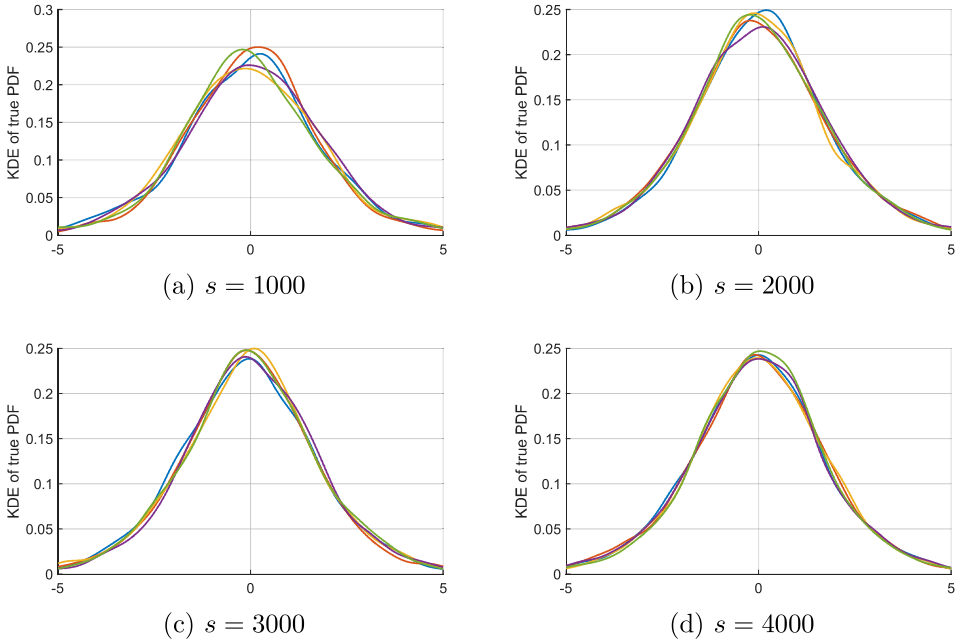
(c) Gamma rozdělení



(d) Weibullovo rozdělení

Obrázek A.4: Zobrazení ECDF p-hodnot pro vybraná rozdělení pozorování a váhy z rozdělení Beta$(4, 2)$. Testy byly provedeny pro dva vážené soubory, každý o velikosti $s = 1000$.

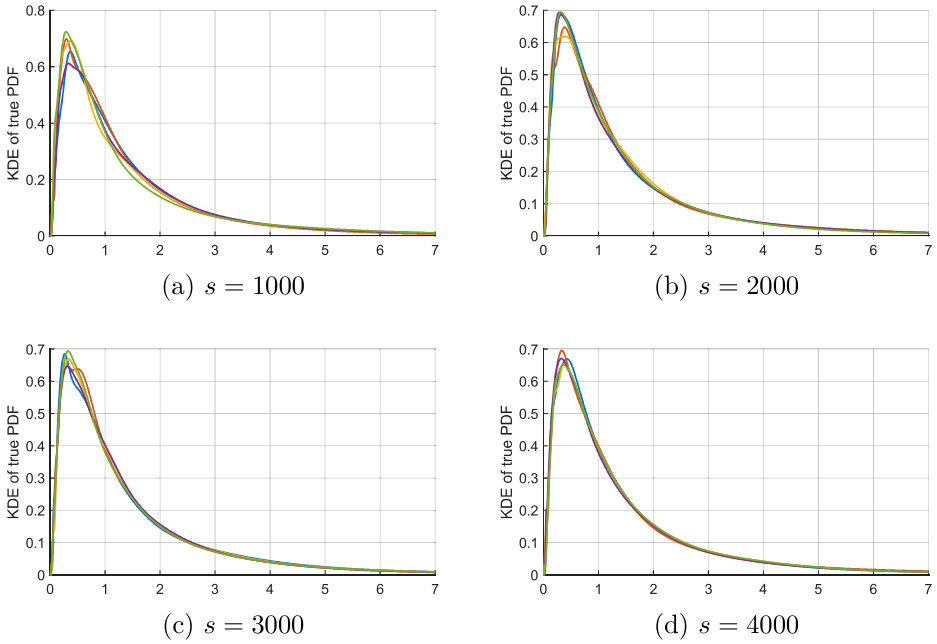
A.3 Opakovaná konstrukce KDE



(a) $s = 1000$ (b) $s = 2000$

(c) $s = 3000$ (d) $s = 4000$

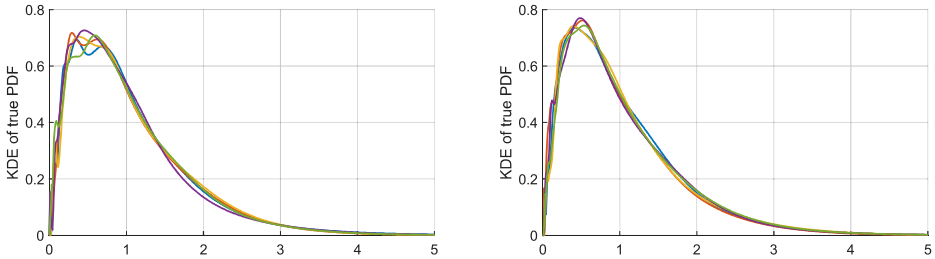
Obrázek A.5: Konstrukce KDE pro pět realizací náhodného výběru s pozorování z Logistic$(0, 1)$.



(a) $s = 1000$ (b) $s = 2000$

(c) $s = 3000$ (d) $s = 4000$

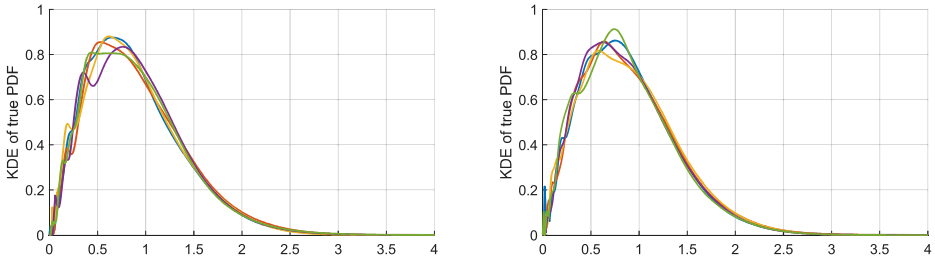
Obrázek A.6: Konstrukce KDE pro pět realizací náhodného výběru s pozorování z Lognormal$(0, 1)$.



(a) $s = 1000$ (b) $s = 2000$

(c) $s = 3000$ (d) $s = 4000$

Obrázek A.7: Konstrukce KDE pro pět realizací náhodného výběru s pozorování z Gamma$(2, 0.5)$.

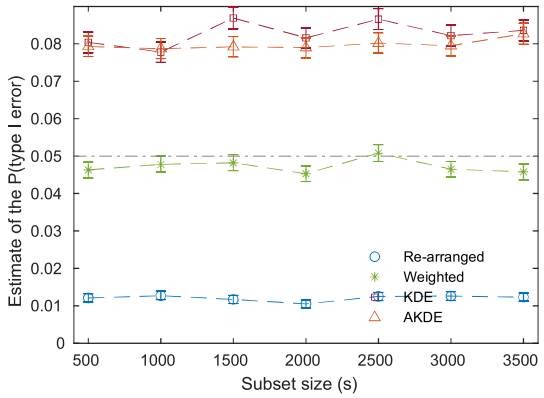


(a) $s = 1000$ (b) $s = 2000$

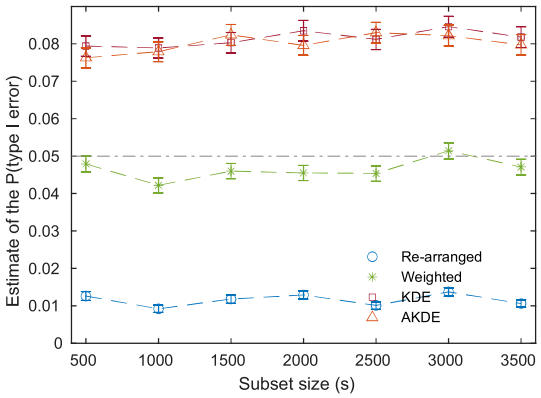
(c) $s = 3000$ (d) $s = 4000$

Obrázek A.8: Konstrukce KDE pro pět realizací náhodného výběru s pozorování z Weibull$(1, 2)$.

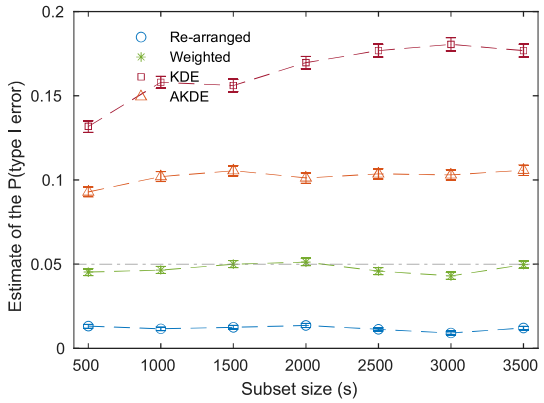
A.4 Testy váženého a neváženého souboru dat



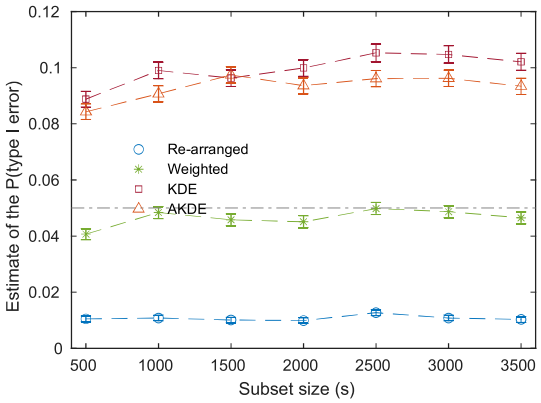
(a) Normální rozdělení



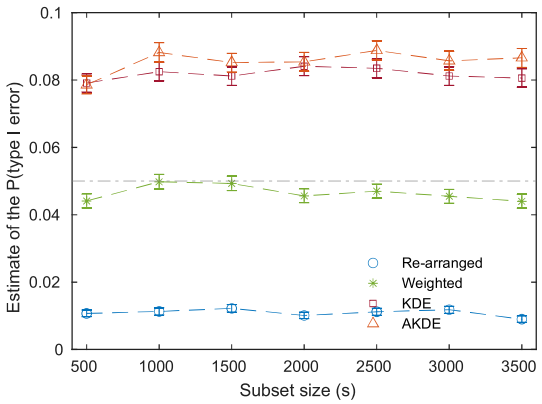
(b) Logistické rozdělení



(c) Lognormalní rozdělení

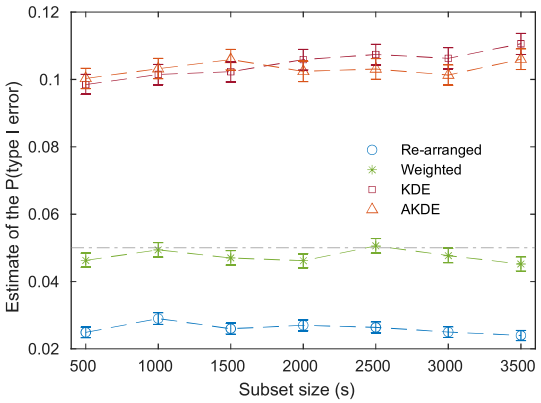


(d) Gamma rozdělení

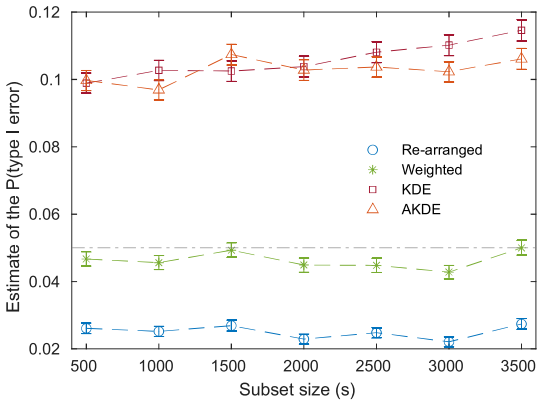


(e) Weibullovo rozdělení

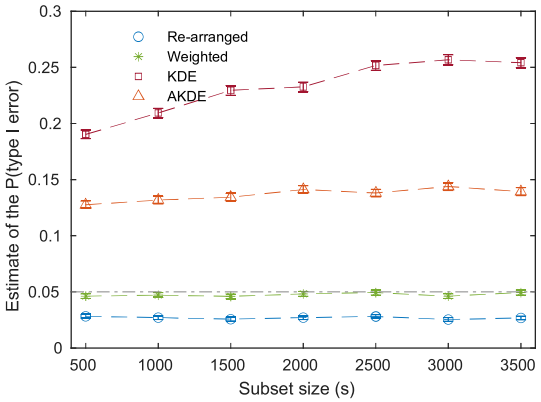
Obrázek A.9: Odhad pravděpodobnosti chyby I. druhu pro testy váženého souboru o s pozorováních s neváženým. Rozdělení pozorování $\boldsymbol{x}$ prvního souboru je uvedeno v popiscích grafů, rozdělení vah $\boldsymbol{w}$ bylo Beta$(2, 4)$.



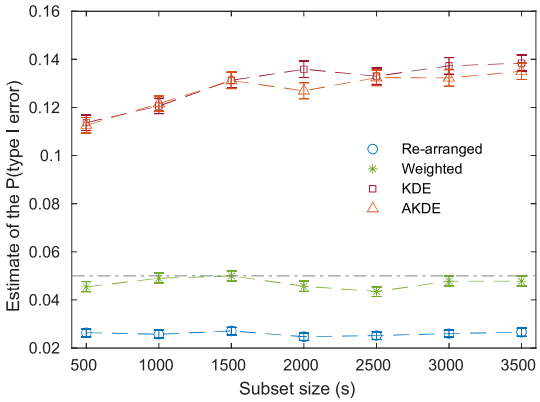
(a) Normální rozdělení



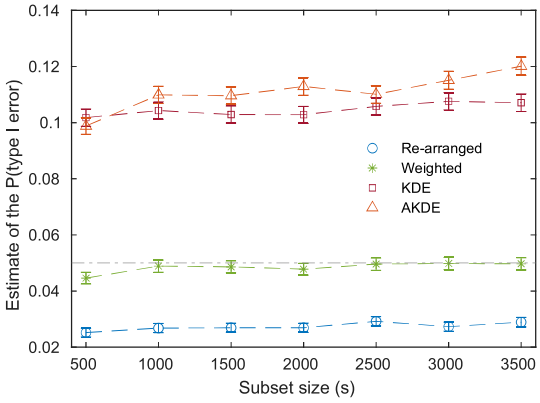
(b) Logistické rozdělení



(c) Lognormalní rozdělení

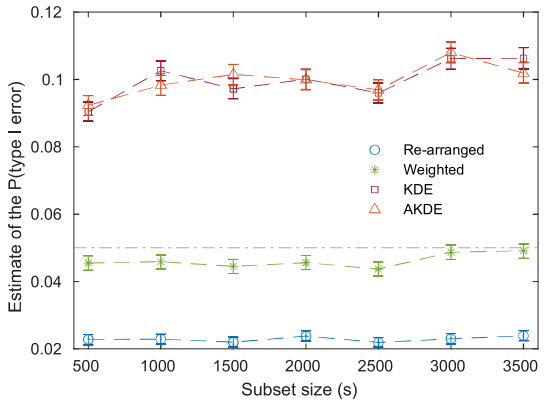


(d) Gamma rozdělení

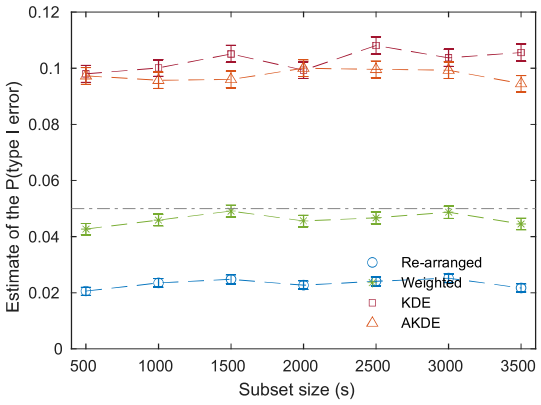


(e) Weibullovo rozdělení

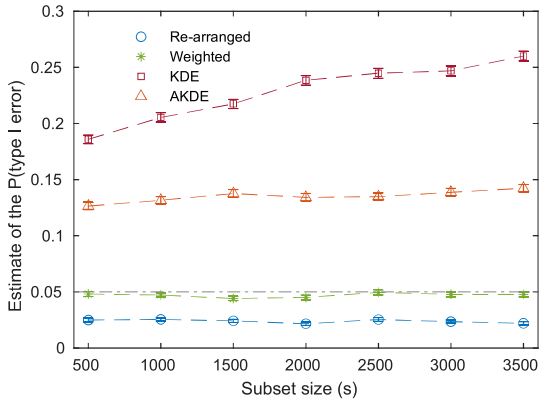
Obrázek A.10: Odhad pravděpodobnosti chyby I. druhu pro testy váženého souboru o s pozorováních s neváženým. Rozdělení pozorování $\boldsymbol{x}$ prvního souboru je uvedeno v popiscích grafů, rozdělení vah $\boldsymbol{w}$ bylo Beta$(0.7, 0.7)$.



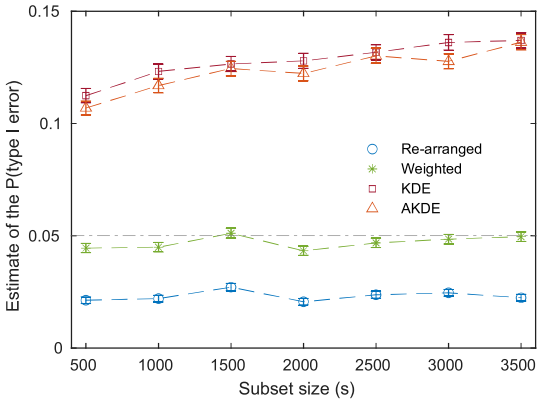
(a) Normální rozdělení



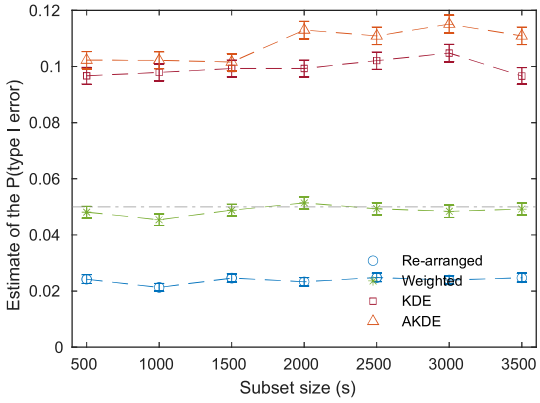
(b) Logistické rozdělení



(c) Lognormalní rozdělení



(d) Gamma rozdělení



(e) Weibullovo rozdělení

Obrázek A.11: Odhad pravděpodobnosti chyby I. druhu pro testy váženého souboru o s pozorováních s neváženou množinou pozorování. Rozdělení pozorování $\boldsymbol{x}$ prvního souboru je uvedeno v popiscích grafů, rozdělení vah $\boldsymbol{w}$ bylo $\mathrm{U}(0, 1)$.