Onderwerpen Masterproef Opleiding Wiskunde
Academiejaar 2021-2022
Vakgroep Toegepaste Wiskunde, Informatica en Statistiek

Titel: Veralgemeende Fourier transformaties: discreet versus continu
Promotoren: prof. Hendrik De Bie (Hendrik.DeBie@UGent.be); prof. Joris Van der Jeugt (Joris.VanderJeugt@UGent.be)

Korte beschrijving:
De klassieke Fourier transformatie (FT) kan men linken aan een realisatie in termen van differentiaaloperatoren van de Lie algebra sl_2. Deze herschrijving laat vervolgens toe om vergaande veralgemeningen van de FT te bekomen, door nieuwe differentiaaloperator realisaties van sl_2 (of zelfs ingewikkeldere Lie (super)algebra's) te construeren. De laatste jaren vormt dit een actief domein van onderzoek, waar zowel continue als discrete transformaties ingevoerd werden. In veel gevallen is er bovendien een interessante kwantummechanische interpretatie.
Doel van deze scriptie is om een aantal dergelijke transformaties in detail te bestuderen, en in het bijzonder een vergelijkende studie te maken van het discrete en continue geval.

Titel: Constant term identities
Promotor: prof. Joris Van der Jeugt (Joris.VanderJeugt@UGent.be)

Korte beschrijving:
In de literatuur zijn er tal van intrigerende "constant term identities" te vinden, die dikwijls hun oorsprong vinden in algebra of combinatoriek. Het berekenen of bepalen van de constante term van een rationale functie (van een bepaalde klasse) is niet eenvoudig. Meestal moet men zich wenden tot technieken die dan kunnen geïmplementeerd worden in computeralgebrapakketten.
Voor deze scriptie onderzoekt de student eerst een gebied waarin de berekening van constante termen van belang is, zoals de bepaling van Kronecker coëfficiënten voor symmetrische functies. Hiervoor is kennis van symmetrische functies en karakters van S_n een pluspunt.
Vervolgens onderzoekt de student enkele methodes/algoritmes om constante term berekeningen te doen. Implementatie van zulke algoritmes in Maple (of Sage) maakt ook deel uit van het werk.
Referentiewerken:
- A. Garsia, N. Wallach, G. Xin and M. Zabrocki, Kronecker coefficients via symmetric functions and constant term identities. (zie o.a. arXiv:0810.0060 [math.CO]).
- G. Xin, A fast algorithm for MacMahon's partition analysis (arXiv:math/0408377 [math.CO]).

Titel: Afhankelijke kansvariabelen en hun som
Promotor: prof. David Vyncke (david.vyncke@ugent.be)

Korte beschrijving:
Het risico waaraan een financiële of verzekeringsportefeuille blootgesteld is, hangt nauw samen met de verdeling van een som van kansvariabelen. Indien de kansvariabelen onafhankelijk zijn, kan men die som schrijven als een convolutie, maar een dergelijke voorwaarde is in de praktijk zelden voldaan. Rekening houden met de reële afhankelijkheidsstructuur brengt echter heel wat problemen met zich mee. Arbenz et al (2011) ontwierpen een algoritme dat de verdeling snel zou moeten berekenen,

maar dat algoritme is beperkt tot positieve kansvariabelen en vereist bovendien dat de volledige copula gekend is. In de praktijk is echter vaak slechts gedeeltelijke informatie over de afhankelijkheid bekend. Door gebruik te maken van dergelijke informatie is het mogelijk om de (verdeling van de) som te begrenzen, zoals geïllustreerd in Bernard et al (2017) en Lux & Papapantoleon (2019). In deze masterproef bestudeert de student verscheidene technieken om de verdeling van een som van afhankelijke kansvariabelen te berekenen en/of te begrenzen.

Referenties:

- Arbenz P., Embrechts P. & Puccetti G. (2011). The AEP algorithm for the fast computation of the distribution of the sum of dependent random variables. Bernoulli 17(2), 562–591.
- Bernard C., Rüschendorf L. & Vanduffel S. (2017). Value-at-risk bounds with variance constraints. Journal of Risk and Insurance 84, 923–959.
- Lux T. & Papapantoleon A. (2019). Model-free bounds on Value-at-Risk using extreme value information and statistical distances. Insurance: Mathematics and Economics 86, 73-83.

Titel: Een ander onderwerp in de financiële of actuariële wiskunde

Promotor: prof. David Vyncke (david.vyncke@ugent.be)

Korte beschrijving:

Bespreekbaar

Titel: Alles normaal of niet ?

Promotor: prof. Christophe Ley (christophe.ley@ugent.be)

Korte beschrijving:

De normale verdeling heeft een speciale plaats in kansrekening, statistiek en, algemeen, in wetenschappen. Dit heeft historische en wiskundige redenen. Er zijn heel wat verschillende stellingen die zeggen dat slechts de normale verdeling een speciale eigenschap bezit: maximale entropie, limiet van een som van toevalsveranderlijken, onafhankelijkheid tussen positie en dispersie schatters, etc.

Het doel van deze masterproef is een studie van de verschillende stellingen die de normale verdeling karakteriseren, en een vooruitzicht op beperkingen van de normale verdeling.

Titel: Information theory

Promotor: prof. Christophe Ley (christophe.ley@ugent.be)

Korte beschrijving:

De historische werken van Claude Shannon over transport van informatie vormen de basis voor een hoog interessant onderzoeksdomein: information theory. Er zijn verschillende maten van informatie: entropie, wederzijds informatie, Kullback-Leibler divergence, etc. Deze maten zijn verbonden met codetheorie.

Over de laatste jaren hebben onderzoekers meer en meer links gevonden tussen information theory en statistiek. Het doel van deze masterproef is het bestuderen van deze links en dus van een gloednieuw domein.

Referentiewerken:
er zijn veel goede tekstboeken.
Voorbeeld: Cover, T. and Thomas, J. (2006) : Elements of Information Theory. Wiley New York.

Titel: Hedgen van basisrisico met behulp van reinforcement learning
Promotor: Prof. dr. M. Vanmaele (michele.vanmaele@ugent.be)
Doelgroep: studenten uit de tweede master Wiskunde

Korte beschrijving:
Basisrisico wordt gedefinieerd als het inherente risico dat een handelaar neemt bij het afdekken van een positie door een tegengestelde positie te nemen in een derivaat van het actief, zoals een futurescontract. Dit basisrisico wordt aanvaard in een poging het prijsrisico af te dekken. Bijvoorbeeld, als de huidige spotprijs, of contante marktprijs, van goud 1190 EUR is, en de prijs van goud in het goudtermijncontract in juni 1195 EUR bedraagt, is de basis, het verschil, 5,00 EUR. Het basisrisico is het risico dat de prijs van de futures niet beweegt in normale, gestage correlatie met de prijs van de onderliggende waarde en dat deze fluctuatie in de basis de effectiviteit van een afdekkingsstrategie (die wordt gebruikt om de blootstelling van een handelaar aan mogelijk verlies te minimaliseren) teniet kan doen. Het verschil tussen de contante prijs en de prijs van de futures kan ofwel groter of kleiner worden tussen het moment waarop een afdekkingspositie wordt ingenomen en het moment waarop die positie wordt gesloten.
Er zijn verschillende types basisrisico te onderscheiden namelijk basisrisico verbonden aan de prijs, de locatie, de kalender of de kwaliteit van het product.
De bedoeling van deze masterscriptie is om hedgingstrategieën te bestuderen in de aanwezig-heid van basisrisico en deze te vergelijken met een hedgingstrategie gebaseerd op reinforce-ment learning. Reinforcement learning (RL) RL is een gebied van machine learning dat zich bezighoudt met hoe intelligente actoren acties moeten ondernemen in een omgeving om een cumulatieve beloning te maximaliseren. Hiervoor kan er voortgebouwd worden op een masterproef van 2020-2021 waar een RL algoritme op basis van "deep determinist c policy gradient" is gebruikt.
Relevante literatuur vormen de volgende artikels:

- Monoyios, M. (2004). Performance of utility-based strategies for hedging basis risk. Quantitative Finance, 4(3), 245–255.
- Trottier, D. A., Godin, F., & Hamel, E. (2018). Local hedging of variable annuities in the presence of basis risk. ASTIN Bulletin, 48(2), 611–646.
- Axel F.A. Adam, Ingmar Nolte (2011). Cross hedging under multiplicative basis risk. Journal of Banking & Finance, 35:2956–2964.
- Broll, U., Welzel, P. & Wong, K. P. (2015) Futures hedging with basis risk and expectation dependence. International Review of Economics, 62(3):213–221.
- Sutton, R. & Barto, A.G. (2018) Reinforcement learning: An introduction, MIT press
- Xue, X., Zhang, J. & Weng, C. (2019) Mean-variance hedging with basis risk. Applied Stochastic Models in Business and Industry, 35:704–716

Titel: Het prijzen en hedgen van kwetsbare opties
Promotor: Prof. dr. M. Vanmaele (michele.vanmaele@ugent.be)
Doelgroep: studenten uit de tweede master Wiskunde

Korte beschrijving:
Bij de bepaling van de prijs van een financieel afgeleid product wordt er doorgaans ondersteld dat de uitgevende instelling financieel gezond is. Onder deze onderstelling mag het falingsrisico van de uitgever verwaarloosd worden. In realiteit is falingsrisico aanwezig ook bij heel grote bedrijven of instellingen zij het eerder in de zogenaamde over-the-counter (OTC)

markt omdat de handel daar niet zo georganiseerd en gereglementeerd is als in het geval er een clearinghouse actief is. (Een clearinghouse zorgt voor de administratie van transacties, saldeert deze, houdt margeverplichtingen bij en garandeert de uiteindelijke betaling en levering van stukken zoals overeengekomen in de transactie.) Vandaar dat houders van dergelijke contracten kwetsbaar zijn voor falingsrisico. Kwetsbare opties werden voor het eerst bestudeerd door Johnson en Stulz (1987). Sindsdien is het prijzen en hedgen van dergelijke opties onder verschillende modellen voor de onderliggende in de literatuur uitvoerig bestudeerd. De bedoeling van deze masterproef is om bepaalde types (bijv. Europees, Aziatisch) kwetsbare opties voor bepaalde types van modellen (bijv. met stochastische volatiliteit of met sprongen) voor de onderliggende te bestuderen. Voor deze studie kan onder meer gesteund worden op volgende artikels en hun referenties:

- Johnson, H., and R. Stulz. (1987). "The pricing of options with default risk". Journal of Finance, 267–280.
- Hui, C., C. Lo, and H. Lee. (2003). "Pricing Vulnerable Black-Scholes Options with Dynamics Default Barriers." Journal of Derivatives 10: 62–69. doi:10.3905/jod.2003.319206.
- Jeon, J., J. Yoon, and M. Kang. (2016). "Valuing Vulnerable Geometric Asian Options." Computers and Mathematics with Applications 71: 676–691. doi:10.1016/j.camwa.2015.12.038.
- Pasricha, P. and A. Goel. (2019). "Pricing vulnerable power exchange options in an intensity based framework", Journal of Computational and Applied Mathematics, Volume 355, 2019, Pages 106-115, https://doi.org/10.1016/j.cam.2019.01.019.
- Yang, S., M. Lee, and J. Kim. (2014). "Pricing Vulnerable Options under a Stochastic Volatility Model." Applied Mathematics Letters 34: 7–12. doi:10.1016/j.aml.2014.03.007.
- Wang, X. (2021). "Analytical valuation of vulnerable European and Asian options in intensity-based models", Journal of Computational and Applied Mathematics 393: 113412, https://doi.org/10.1016/j.cam.2021.113412

Titel: Stein's lemma en toepassingen in portefeuilleselectieproblemen
Promotor: Prof. dr. M. Vanmaele
Doelgroep: studenten uit de tweede master Wiskunde

Korte beschrijving:
Voor een bivariate normaalverdeelde vector $(X , Y)^T$, heeft Stein (1973, 1981) aangetoond dat
Cov [h (X) , Y] = Cov [X, Y] E [h ' (X)]voor een willekeurige afleidbare functie h waarvoor E[| h '(X)|] bestaat. Dit lemma heeft zijn nut bewezen in heel wat disciplines, zoals statistiek, waarschijnlijkheid, besliskunde en financiën. Rendementen van aandelen zijn echter niet altijd symmetrisch verdeeld, ze vertonen scheefheid. Deze observatie zette Adcock ertoe aan om in een serie van artikels een type van Stein's lemma op te stellen voor multivariate verdelingen die consistent zijn met Simaan's (1993) modelering van aandelenrendementen. Landsmann en co-auteurs hebben verdere uitbreidingen van Stein's lemma bewezen voor (multivariate) elliptisch verdeelde veranderlijken met toepassingen in o.a. risicotheorie (zie Landsmann & Valdez, 2016). Vanduffel en Yao (2017) vertrekken van Simaan's setting en stellen een nieuw type van Stein's lemma op voor de multivariate veralgemeende hyperbolische verdeling. Als toepassing, beschouwen ze een portefeuilleselectievraagstuk.
De opdracht van deze masterproef bestaat in het bestuderen van Stein's lemma voor verschillende verdelingen en de toepassingen ervan in portefeuilleselectieproblemen en bij risicomaten.

Voor deze studie kan gesteund worden op de volgende artikels:

- C.J. Adcock (2014). Mean–variance–skewness efficient surfaces, Stein's lemma and the multivariate extended skew-Student distribution. European Journal of Operational Research, 234:392–401
- Z. Landsmann and E. Valdez (2016). The tail Stein's identity with applications to risk measures. North American Actuarial Journal, 20(4): 313-326.
- Y. Simaan (1993). Portfolio selection and asset pricing-three parameter framework. Management Science, 39 (5):568–587 .C.M. Stein (1973). Estimation of the mean of a multivariate normal distribution. In Proceedings of the Prague symposium on asymptotic statistics (pp. 345–381) .
- C. M. Stein (1981). Estimation of the mean of a multivariate normal distribution. Annals of Statistics, 9:1135–1151 .

- S. Vanduffel and J. Yao (2017). A Stein type lemma for the multivariate generalized hyperbolic distribution. European Journal of Operational Research, 261(2):606-612

Titel: Filtering, Markov chain Monte Carlo (MCMC) met toepassingen in financiële wiskunde
Promotor: Prof. dr. M. Vanmaele
Doelgroep: studenten uit de tweede master Wiskunde

Korte beschrijving:
Bij het prijzen van financiële en verzekeringsproducten gaat men ervan uit dat het mogelijk is om alle bronnen van onzekerheid te observeren. Dit is een onrealistische onderstelling. Bijvoorbeeld de stochastische volatiliteit van aandelenprijzen waarop opties geschreven zijn, is vaak maar gedeeltelijk waarneembaar. Of een belegger wil een afgeleid product prijzen waarvan de onderliggende zelf niet continu waarneembaar is maar slechts op discrete tijdstippen. Ook al hebben beleggers toegang tot de volledige markt, moeten ze vaak beslissingen nemen op basis van partiële informatie. Bij levensverzekeringsproducten kunnen verzekeraars de individuele sterfte van verzekeringsnemers waarnemen maar niet de zogenaamde "mortality rate". Dus ook hier is er sprake van partiële informatie. Bij het prijzen kan dan gebruik gemaakt worden van filters zoals de Kalman-filter in het geval van een lineair, Gaussisch model. Gaat het echter om een niet-lineair, niet-Gaussisch model dan kan "particle filtering" of "sequential Monte Carlo" gebruikt worden.
Het doel van deze masterproef is het bestuderen van deze filtertechniek(en) en de toepassingen ervan in het bijzonder bij het prijzen van financiële producten of bij portefeuille-optimalisatie.

Voor deze studie kan gesteund worden op volgende artikels:
- Creal, D. (2008). Analysis of filtering and smoothing algorithms for Lévy-driven stochastic volatility models. Computational Statistics & Data Analysis 52:2863–2876.
- Desai, R., Lele, T., Viens, F. (2003). A Monte-Carlo method for portfolio optimization under partially observed stochastic volatility. 2003 IEEE International Conference on Computational Intelligence for Financial Engineering. Proceedings, pp. 257-263.
- Fouqué, J.-P., Papanicolaou, A., Sircar, R. (2015). Filtering and portfolio optimization with stochastic unobserved drift in asset returns. Communications in Mathematical Sciences 13(4): 935–953
- Lindström, E., Ströjby, J., Brodén, M., Wiktorsson, M., Holst, J. (2008). Sequential calibration of options. Computational Statistics & Data Analysis 52:2877–2891.
- Rambharat, B.R. (2012). American option valuation with particle filters. In: R.A. Carmona et al. (eds.), Numerical Methods in Finance, Springer Proceedings in Mathematics 12, pp. 51-82

Titel: Een ander onderwerp in het domein van de financiële of actuariële wiskunde
Promotor: Prof. dr. M. Vanmaele
Doelgroep: studenten uit de tweede master Wiskunde

Bespreekbaar

Titel: Statistische besluitvorming na gebruik van machine learning technieken
Promotor: prof. Stijn Vansteelandt (stijn.vansteelandt@ugent.be)

Korte beschrijving:
Meer en meer statistische analyses maken gebruik van machine learning technieken om bijvoorbeeld het effect van een behandeling of interventie te schatten. Zoals alle statistische analyses, genereren ook machine learning technieken onzekerheid op de resultaten (ten gevolge van steekproefvariatie); omwille van de complexiteit van de technieken wordt dit in de verdere analyse echter meestal genegeerd. Als gevolg hiervan zijn de betrouwbaarheidsintervallen die men bekomt, typisch zwaar vertekend. In deze masterproef zult u betrokken worden in recente ontwikkelingen, alsook onderzoek binnen de onderzoeksgroep van de promotor, om na te gaan hoe men hiermee kan omgaan. De bestudeerde methodes zullen theoretische en/of door middel van simulatiestudies en concrete data-analyses worden geëvalueerd.

Titel: Prespecified statistical analysis
Promotor: prof. Stijn Vansteelandt (stijn.vansteelandt@ugent.be)

Korte beschrijving:
Het kunnen vastleggen van een statistisch analyseplan, alvorens de data verzameld werden, is van cruciaal belang om bewuste of onbewuste bijsturing van de resultaten te vermijden. Het is de norm in experimentele klinische studies, en tevens een vereiste van de regulator (bvb. FDA), maar moeilijk in observationele studies. De reden is dat er in dergelijke studies een nood is om te corrigeren voor confounding, en het is moeilijk om op voorhand vast te leggen volgens welke modellen dat zal gebeuren. In deze thesis zullen we daarom recente ontwikkelingen bestuderen die erop gericht zijn om een analyse volledig op voorhand vast te leggen, en deze toepassen in een concrete context.

Titel: Een ander onderwerp in het domein van de mathematische statistiek en/of toegepaste data-analyse
Promotor: prof. Stijn Vansteelandt (stijn.vansteelandt@ugent.be)

Korte beschrijving:
Bespreekbaar

Titel: Statistische besluitvorming na gebruik van machine learning technieken
Promotor: prof. Stijn Vansteelandt (stijn.vansteelandt@ugent.be)

Korte beschrijving:
Meer en meer statistische analyses maken gebruik van machine learning technieken om bijvoorbeeld het effect van een behandeling of interventie te schatten. Zoals alle statistische analyses, genereren ook machine learning technieken onzekerheid op de resultaten (ten gevolge van steekproefvariatie); omwille van de complexiteit van de technieken wordt dit in de verdere analyse echter meestal genegeerd. Als gevolg hiervan zijn de betrouwbaarheidsintervallen die men bekomt, typisch zwaar vertekend. In deze masterproef zult u betrokken worden in recente ontwikkelingen, alsook onderzoek binnen de onderzoeksgroep van de promotor, om na te gaan hoe men hiermee kan omgaan. De bestudeerde methodes zullen theoretische en/of door middel van simulatiestudies en concrete data-analyses worden geëvalueerd.

Title: Assessment of treatment effect heterogeneity using modern machine learning techniques.
Promotor: prof. Stijn Vansteelandt (Stijn.Vansteelandt@UGent.be); Begeleider: dr. Oliver Dukes (Oliver.Dukes@UGent.be)

Korte beschrijving:
In the era of personalized medicine, there is much interest in understanding how the effect to a given treatment depends on an individual's characteristics. Traditionally this has been done using simple parametric statistical models, but in recent years some proposals have been developed to learn about effect heterogeneity in a data-driven way, based on the advances in machine learning. Surprisingly, several of these proposals also facilitate statistical inference on the conditional treatment effect. In this project, you will review and compare the different frameworks, initially from a theoretical perspective, to understand their strengths and weaknesses. Further comparison can then be made via running simulation studies and conducting data analyses.

Titel: Een vergelijking van algoritmes voor de numerieke oplossing van de twee-dimensionale tijdsonafhankelijke Schrödinger
Promotor: prof Marnix Van Daele (marnix.vandaele@ugent.be); Begeleider: Toon Baeyens (toon.baeyens@ugent.be)

Korte beschrijving:
De Schrödinger-vergelijking is een partiële differentiaalvergelijking die de toestand beschrijft van een kwantummechanisch systeem. Van deze vergelijking, oorspronkelijk opgesteld door de Oostenrijkse natuurkundige Erwin Schrödinger, bestaat een tijdsafhankelijke vorm en een tijdsonafhankelijke vorm.
Bij dit thesisonderwerp focussen we ons op de tijdsonafhankelijke vorm in twee ruimtelijke dimensies. Er bestaan verschillende numerieke standaardtechnieken om dit probleem op te lossen (bvb. matrixmethoden op basis van eindige-differenties of eindige-elementen), maar er bestaan ook meer gespecialiseerde methoden zoals matslise2D, wat ontwikkeld werd door Toon Baeyens in het kader van zijn doctoraatsonderzoek. Hierbij wordt gebruik gemaakt van principes en methoden die gebruikt worden in de matlab toolbox Matslise (cfr. Wiskundige modellering). De bedoeling is een vergelijking te maken van de verschillende methoden.
Startpunt/referenties:
voor deze vergelijkende studie kan je starten met enkele bekende methodes te vergelijken. Het is niet de bedoeling dat je zelf grote algoritmen implementeert.

- Ixaru, L. Gr. "New Numerical Method for the Eigenvalue Problem of the 2D
- Schrödinger Equation." Computer Physics Communications 181, no. 10 (October 1, 2010): 1738–42. https://doi.org/10.1016/j.cpc.2010.06.031.
- Braun, M., S. A. Sofianos, D. G. Papageorgiou,
- E. Lagaris. "An Efficient Chebyshev–Lanczos Method for Obtaining Eigensolutions of the Schrödinger Equation on a Grid." Journal of Computational Physics 126, no. 2 (July 1, 1996): 315–27. https://doi.org/10.1006/jcph.1996.0140.

Titel: De numerieke oplossingen van tijdsafhankelijke Schrödinger-vergelijkingen in één en twee dimensies
Promotor: prof Marnix Van Daele (marnix.vandaele@ugent.be); Begeleider: Toon Baeyens (toon.baeyens@ugent.be)

Korte beschrijving:
De Schrödinger-vergelijking is een partiële differentiaalvergelijking die de toestand beschrijft van een kwantummechanisch systeem. Van deze vergelijking, oorspronkelijk opgesteld door de Oostenrijkse natuurkundige Erwin Schrödinger, bestaat een tijdsafhankelijke vorm en een tijdsonafhankelijke vorm. Bij dit thesisonderwerp focussen we ons op de tijdsafhankelijke vorm in een of twee ruimtelijke dimensies.
Er bestaan verschillende numerieke standaardtechnieken om dit probleem op te lossen (bvb. discretisatie via eindige differentie-methodes voor de ruimtelijke veranderlijke en de method of lines voor de tijdsintegratie), maar er bestaan ook

meer gespecialiseerde methoden. In de onderzoeksgroep Numerieke wiskunde wordt een methode ontwikkeld die gebaseerd is op principes/methoden van Matslise (cfr. Wiskundige modellering).
De bedoeling van dit thesisonderwerp is om met behulp van reeds bestaande methoden voor de numerieke oplossing van het tijdsonafhankelijke probleem een nieuwe implementatie te bouwen om ook het tijdsafhankelijk probleem op te lossen.
Startpunt/referenties:

- je kan vertrekken vanaf de geziene technieken in 'Gevorderde numerieke methoden' om gewone differentiaalvergelijkingen op te lossen. Je zal zelf een implementatie bouwen in Sage/Python (of indien je dat wenst C++).
- Ledoux, V., and M. Van Daele. "The Accurate Numerical Solution of the Schrödinger Equation with an Explicitly Time-Dependent Hamiltonian." Computer Physics Communications 185, no. 6 (June 1, 2014): 1589–94. https://doi.org/10.1016/j.cpc.2014.02.023.

Titel: De numerieke oplossing van het 3D Schrödinger-probleem
Promotor: prof Marnix Van Daele (marnix.vandaele@ugent.be); Begeleider: Toon Baeyens (toon.baeyens@ugent.be)

De Schrödingervergelijking is een lineaire tweede orde partiële differentiaalvergelijking, ze kent haar oorsprong in de kwantummechanica. Ze beschrijft de onderliggende golfvergelijking van een kwantummechanisch systeem. Zoals de naam `golfvergelijking' doet vermoeden, zijn oplossingen van de Schrödingervergelijking zeer oscillatorisch. Dit geeft numerieke benaderingsmethoden een extra uitdaging.
In deze masterproef zoeken we numerieke oplossingen van het drie-dimensionale tijdsonafhankelijk Schrödingerprobleem $-\Delta \psi + V(x,y,z) \psi = E \psi$ over (een deel van) R^3. Aan de rand van dit gebied veronderstellen we dat $\psi(x,y,z)= 0$. Dit is een eigenwaarde- en eigenfunctieprobleem.
De techniek die we zullen gebruiken om dit probleem op te lossen, is een uitbreiding van een techniek die we momenteel aan het optimaliseren zijn voor het equivalente twee-dimensionale probleem. Bij deze techniek gebruiken we een discretisatie in twee dimensies. Voor de overige dimensie stellen we een goed gekozen basis op, gebaseerd op de ééndimensionale Schrödinger-vergelijking. Er bestaan, dankzij Matslise, zeer snelle routines om deze basis te construeren.
Om de tweede orde afgeleide te bepalen in de gediscretiseerde assen maken we gebruik van klassieke eindige differentieformules, zoals $h^2 f''(x) = f(x-h) - 2f(x) + f(x+h)$. Om het oscillatorisch gedrag van de eigenfuncties nauwkeurig te benaderen zullen we hoge orde eindige differentie-formules gebruiken.
In deze masterproef zal de student

-zichzelf vertrouwd maken met deze methode in twee dimensies.
- de uitdaging identificeren en oplossingen ervoor voorstellen.
- een implementatie bouwen van deze methode.
- numerieke experimenten uitvoeren om de sterktes en zwaktes van deze techniek te illustreren.

Afhankelijk van de interesses van de studenten kunnen we ons toespitsen op ofwel de wiskundige achtergrond, ofwel een efficiënte implementatie. Toon Baeyens zal instaan voor de begeleiding.

Titel: Topologische optimalisatie van algemene filtraties op simpliciale complexen
Begeleiders(s): Promotor: Yvan Saeys (yvan.saeys@ugent.be); Begeleider: Robin Vandaele (robin.vandaele@ugent.be)

Korte beschrijving:
Persistente homologie is een toepassing uit de algebraïsche topologie, die ons toestaat om onderliggende topologische informatie uit data te extraheren. Binnen machinaal leren wordt persistente homologie zowel toegepast voor gesuperviseerd als niet-gesuperviseerd leren.

Recent onderzoek toont aan dat persistente homologie daarboven ook als optimalisatietechniek kan gebruikt worden om puntenwolken (data) te visualiseren op een manier die gegeven topologische informatie respecteert. Dit wordt gedaan aan de hand van de volgende iteratieve stappen:

1. Een simpliciaal complex (hetgeen kan gezien worden als een veralgemening van een graaf) wordt gebouwd op de huidige visualisatie.
2. De afstandsfunctie op de paren van punten wordt gebruikt om een filtratie te bouwen op het complex, dit zijnde een eindige rij van deelcomplexen geordend volgens inclusie.
3. Persistente homologie wordt berekend van deze filtratie en wordt vervolgens een reële te optimaliseren waarde toegekend.

In dit thesisonderwerp zijn we geïnteresseerd in het veralgemenen van deze methode. Voor puntenwolken beperkt de combinatie van stappen 1 en 2 zich momenteel immers tot twee specifieke filtraties: de Vietoris-Rips- en alpha-filtratie. Dit staat ons toe om de data te visualiseren volgens een aantal clusters of gaten, maar niet volgens andere topologische informatie, zoals bifurcaties.
De student is dus verantwoordelijk voor de uitbreiding van topologische optimalisatie van puntenwolken aan de hand van meer algemene simpliciale complexen (bv. nabijheidsgrafen), alsook de integratie van meer algemene functies die de filtratie definiëren. Meer concreet houdt dit in: de nodige literatuurstudie over topologische data-analyse, de theoretische uitwerking voor het veralgemenen van de optimalisatie via subafgeleiden, het implementeren van de methode of aanpassen van de bestaande methoden hiertoe, alsook de uitwerking van finale experimenten en toepassingen.

Titel: Metric learning for domain adaptive segmentation:
Promotor: Yvan Saeys (yvan.saeys@ugent.be); Begeleider: Joris Roels (jorisb.roels@ugent.be)

Korte beschrijving:
Image segmentation is the process of classifying each pixel of an image in a particular class. To improve generalization, it is desired that the developed classifiers are easily adaptable to new image domains: e.g. a segmentation algorithm that works fine for a particular dataset should not require lots of adjustments to work on a new dataset. This problem is also called domain adaptive segmentation and is far from straightforward with the state-of-the-art approaches.
In this thesis, we will use metric learning approaches to find similarities and differences between different domains and adjust the classifier accordingly to generalize to new domains with as little supervision as possible. To do this, the student will start with a literature study in domain adaptation and metric learning, experiment with deep learning frameworks (e.g. Pytorch or Tensorflow) and develop a new domain adaptive segmentation technique for large-scale biological datasets. The student can make use of a large annotated database of 3D microscopy datasets for training and validation.

Titel: Een studie naar kwantitatieve maten voor het beoordelen van bias en fairness van machine learning modellen
Promotor: Yvan Saeys (yvan.saeys@ugent.be) Begeleider: Arne Gevaert (arne.gevaert@ugent.be)

Korte beschrijving:
Ethische aspecten van Machine learning worden meer en meer belangrijk nu deze methoden in steeds meer aspecten van het dagelijkse leven gebruikt worden. Recent onderzoek heeft echter aangetoond dat machine learning modellen vaak subject kunnen zijn van impliciete biases. Deze kunnen al dan niet vrijwillig in het model geintroduceerd geweest zijn, of een neveneffect zijn van de manier waarop de data verwerkt is. Het is echter belangrijk data deze biases bekend zijn, en dus op een objectieve manier gedecteerd worden. De volgende figuur toont een voorbeeld van biases die aanwezig kunnen zijn in gezichtsherkenningssystemen voor verschillende huidtypes.

Het is belangrijk dat deze biases objectief kunnen gedetecteerd worden, en dat er eventueel ook een onderscheid kan gemaakt worden tussen de verschillende types van biases die kunnen optreden.

In deze masterproef zal de student in een eerste fase een overzicht maken van de verschillende types van biases die kunnen optreden in machine learning modellen. Vervolgens zal er bekeken worden welke objectieve maten opgesteld kunnen worden om biases in machine learning modellen te kwantificeren. De student zal ook zelf experimenten kunnen opstellen waarbij verschillende biases in datasets worden geintroduceerd, om zo te gaan valideren welke maten tot een goede detectie van biases leiden, en zo tot nieuwe guidelines te komen die in de toekomst kunnen leiden tot meer ethisch verantwoorde machine learning modellen.

Titel: Een ander onderwerp in het domein van machinaal leren
Promotor: Yvan Saeys (yvan.saeys@ugent.be)

Korte inhoud: bespreekbaar met geïnteresseerde studenten

Titel: Bias en variabiliteit in de data-analyse van genexpressiestudies met herhaalde metingen
Promotoren: prof. Lieven Clement (lieven.clement@ugent.be); prof. Stijn Vansteelandt (stijn.vansteelandt@ugent.be)

Korte beschrijving:
Dankzij hoge-doorvoer sequeneringstechnologie kan genexpressie (RNA-seq) simultaan worden geëvalueerd op het niveau van duizenden individuele transcripten. De data in RNA-seq studies zijn telprofielen waarbij de tellingen voor een bepaald gen een proxy zijn voor de concentratie in een staal.
De teldata worden typisch gemodelleerd aan de hand van negatief binomiale (NB) modellen aangezien de tellingen in biologische studies overdispers zijn t.o.v. een Poisson distributie omwille van de biologische variabiliteit. In de literatuur werden reeds flexibele software-tools ontwikkeld voor het modelleren van data afkomstig van complexe studiedesigns waarbij de parallelle datastructuur van genomische studies wordt uitgebuit om betere schattingen te bekomen van variantie parameters door informatie te ontlenen over de genen heen. De huidige RNA-seq tools zijn echter niet ontwikkeld om correlatiestructuren te modelleren in experimenten met herhaalde metingen en longitudinale designs.
In de RNA-seq literatuur worden subject-specifieke effecten vaak gemodelleerd door het introduceren van additionele vaste effecten in de lineaire predictor van het NB model. In de econometrische literatuur is echter beschreven dat het gebruik van de volledige set van individuele dummy variabelen in NB modellen kan leiden tot inconsistente schatters voor de overige vaste factoren. Daarenboven modelleert de gerandomiseerde compleet blok (RCB) analyse in conventionele RNA-seq tools enkel de binnen-subject error. Hierdoor kunnen deze tools sowieso niet worden ingezet voor inferentie over effecten die zich manifesteren tussen subjecten aangezien de huidige implementatie van de statistische toetsen de variabiliteit tussen subjecten niet in rekening kan brengen.
Het doel van deze masterproef is om inzicht te krijgen in het negatief binomiaal regressiemodel. Hierbij zal eerst de bias en de controle van type I fouten worden geëvalueerd van de RCB analyse met standaard RNA-seq tools waarbij zowel effecten binnen als tussen subjecten zullen worden bestudeerd. We zullen gebruik maken van zowel gesimuleerde alsook echte RNA-seq data. In het tweede deel van de thesis zal worden geëxploreerd hoe kan worden gecorrigeerd voor de eventuele bias en voor de variabiliteit binnen en tussen subjecten in de statistische toetsen die worden gebruikt in RNA-seq studies.
Dankzij hoge-doorvoer sequeneringstechnologie kan genexpressie (RNA-seq) simultaan worden geëvalueerd op het niveau van duizenden individuele transcripten. De data in RNA-seq studies zijn telprofielen waarbij de tellingen voor een bepaald gen een proxy zijn voor de concentratie in een staal.
De teldata worden typisch gemodelleerd aan de hand van negatief binomiale (NB) modellen aangezien de tellingen in biologische studies overdispers zijn t.o.v. een Poisson distributie omwille van de biologische variabiliteit. In de literatuur werden reeds flexibele software-tools ontwikkeld voor het modelleren van data afkomstig van complexe studiedesigns waarbij de

parallelle datastructuur van genomische studies wordt uitgebuit om betere schattingen te bekomen van variantie parameters door informatie te ontlenen over de genen heen. De huidige RNA-seq tools zijn echter niet ontwikkeld om correlatiestructuren te modelleren in experimenten met herhaalde metingen en longitudinale designs.
In de RNA-seq literatuur worden subject-specifieke effecten vaak gemodelleerd door het introduceren van additionele vaste effecten in de lineaire predictor van het NB model. In de econometrische literatuur is echter beschreven dat het gebruik van de volledige set van individuele dummy variabelen in NB modellen kan leiden tot inconsistente schatters voor de overige vaste factoren. Daarenboven modelleert de gerandomiseerde compleet blok (RCB) analyse in conventionele RNA-seq tools enkel de binnen-subject error. Hierdoor kunnen deze tools sowieso niet worden ingezet voor inferentie over effecten die zich manifesteren tussen subjecten aangezien de huidige implementatie van de statistische toetsen de variabiliteit tussen subjecten niet in rekening kan brengen.

Het doel van deze masterproef is om inzicht te krijgen in het negatief binomiaal regressiemodel. Hierbij zal eerst de bias en de controle van type I fouten worden geëvalueerd van de RCB analyse met standaard RNA-seq tools waarbij zowel effecten binnen als tussen subjecten zullen worden bestudeerd. We zullen gebruik maken van zowel gesimuleerde alsook echte RNA-seq data. In het tweede deel van de thesis zal worden geëxploreerd hoe kan worden gecorrigeerd voor de eventuele bias en voor de variabiliteit binnen en tussen subjecten in de statistische toetsen die worden gebruikt in RNA-seq studies.

Titel: Splines voor het schatten van niet-lineaire trends in longitudinale proteomics experimenten
Promotoren: prof. Lieven Clement (lieven.clement@ugent.be); prof. Lennart Martens (lennart.martens@ugent.be)

Korte beschrijving:
Vele biologische processen zijn dynamisch en tijdreeks 'omics experimenten waarbij de expressie van duizenden genen of proteines worden gevolgd over de tijd zijn essentieel om deze processen verder te ontrafelen. De analyse van high-throughput tijdreeks data is erg uitdagend gezien de temporele effecten van stimuli op de expressie van proteines typisch niet-lineair zijn. Daarom is regularisatie nodig om het onderliggende continue signaal te schatten uit de geobserveerde expressie profielen die typisch sterk variabel zijn. Splines zijn populair binnen de context van 'omics tijdreeks experimenten. Ze kunnen niet-lineaire temporele expressiepatronen schatten op een datagedreven manier, maar splines zijn nog niet goed geïntroduceerd in het veld van de label-vrije differentiele proteomics data-analyse.
In massa-spectrometrie gebaseerde proteomics neemt de kwatificatie plaats op het niveau van de peptides (stukjes van de proteines) en de peptide intensiteiten moeten worden geaggregeerd op het niveau van de proteinen. Recent hebben Goeminne et al. (2016, doi: 10.1074/mcp.M115.055897) MsqRob ontwikkeld, een robust lineair model voor het detecteren van differentiële abundante proteinen dat onmiddellijk peptide intensiteiten modelleert. De methode aggregeert de peptide data binnen het model en heeft een hogere sensitiviteit en specificiteit dan de hangbare methoden in de literatuur. MSqRob laat echter nog niet toe om niet-lineaire trends te schatten in tijdreeks data. De software-tool is ingebed binnen het raamwerk van gemengde modellen waarbij de data kan worden gemodelleerd met vaste effecten (temporele trend, effect van de stimulus,...) en random effecten (sample en subject-specifieke effecten).
Het doel van deze masterproef is (a) om inzicht te krijgen in schattings- en inferentie methoden voor splines, (b) hoe splines binnen het raamwerk van gemengde modellen kunnen worden ingepast, (c) om de methode van Goeminne et al. uit te breiden voor de analyse van longitudinale studies waarbij het proteoom van patiënten of celculturen wordt gevolgd over de tijd en (d) om de uitbreiding te implementeren binnen het software pakket MSqRob (https://github.com/statOmics/MsqRob). Hierbij ligt de nadruk zowel op het schatten van niet-lineaire trends binnen een biologische conditie en om verschillen in temporele expressie-profielen op te pikken tussen biologische condities (e.g. ziek vs gezond, behandeld vs onbehandeld, ...). Het uitgangspunt is het boek "Semiparametric regression" van D. Ruppert, M. Wand en R. Carrol, Cambridge series in statistical and probablistic mathematics (H1-H6).

Titel: Een ander onderwerp in het domein van de statistische bioinformatica en/of toegepaste data-analyse
Promotor: prof. Lieven Clement (lieven.clement@ugent.be)

Korte beschrijving: Bespreekbaar

Titel: Transitiviteit van vaagrelaties voor het modelleren van benaderende gelijkheid met toepassing in machine learning.
Promotor: prof. Chris Cornelis (chris.cornelis@ugent.be)

Korte beschrijving
De klassieke gelijkheidsrelatie in een universum U is een equivalentierelatie: inderdaad, elk element van U is gelijk aan zichzelf (reflexief), als x gelijk is aan y, is y ook gelijk aan x (symmetrisch) en uit de gelijkheid van x en y samen met die van y en z volgt automatisch dat x en z gelijk zijn (transitief).
Als we toelaten dat elementen op elkaar kunnen lijken in een bepaalde mate, spreken we van benaderende gelijkheid. In deze masterproef modelleren we dit a.d.h.v. een vaagrelatie R in U, d.w.z. een afbeelding van U^2 naar [0,1]. Meestal voeren we enkele beperkingen in voor de relatie R, zoals:

- $R(x,x) = 1$ (reflexief)
- $R(x,y) = R(y,x)$ (symmetrisch)
- $R(x,z) \geq \min(R(x,y),R(y,z))$ (min-transitief)

Ook varianten hierop zijn mogelijk, zo kan men bijvoorbeeld het minimum in de transitiviteitsvoorwaarde vervangen door het product of door een andere operator. Soms laat men deze voorwaarde zelfs gewoon weg, vanuit de gedachte dat benaderende gelijkheid in het algemeen niet transitief is (wat reeds werd opgemerkt door de bekende wiskundige Henri Poincaré).
Het doel van deze masterproef bestaat erin om in te zoomen op de transitiviteitsvoorwaarde. Afhankelijk van de interesse van de student kan hierbij een meer theoretische richting gevolgd worden door na te gaan:

- welke verschillende vormen van afgezwakte transitiviteit werden voorgesteld in de literatuur
- welke transitiviteitseigenschappen voldaan zijn voor de meest courante benaderende gelijkheidsmaten
- welke constructiemethodes er bestaan voor benaderende gelijkheidsmaten met gegeven transitiviteitseigenschappen, of er kan ook worden toegespitst op machine learning, waarbij het de bedoeling is om na te gaan via een experimentele studie of het opleggen van afgezwakte transitiviteit de nauwkeurigheid van similariteitsgebaseerde classificatie-algoritmen al of niet kan verbeteren.
  Er kan voor deze opdracht vertrokken worden van een eerdere masterproef die het onderwerp binnen een ruimer kader bestudeerde.

Titel: Reductie in coveringgebaseerde ruwverzamelingenleer: een vergelijkende studie
Promotor: prof. Chris Cornelis (chris.cornelis@ugent.be)

Korte beschrijving:
Ruwverzamelingenleer (Eng., rough set theory) is een wiskundige theorie waarbij men deelverzamelingen A van een universum U benadert m.b.v. een equivalentierelatie R over U, of gelijkwaardig hiermee, een partitie P van U. Meer bepaald omvat de onderbenadering van A alle equivalentieklassen van R die volledig bevat zitten in A en de bovenbenadering van A alle klassen die een niet-ledige doorsnede hebben met A. Deze theorie kent uitgebreide toepassing binnen data-analyse, waar ze onder meer gebruikt wordt voor de reductie van classificatietabellen door het selecteren van de meest informatieve objectkenmerken.

Wanneer men de eis dat P een partitie van U vormt minder strikt maakt door slechts te eisen dat P een bedekking of covering vormt, spreekt men over coveringgebaseerde ruwverzamelingen. De definitie van onder- en bovenbenadering is in dit geval niet langer eenduidig bepaald. Zo werden er in de literatuur tientallen verschillende paren van benaderingsoperatoren gebaseerd op een covering voorgesteld. De studie van de verbanden tussen verschillende definities en de eigenschappen die ze vervullen is een relevant en actueel onderzoeksthema.
In deze masterproef dient de student de beschikbare literatuur over reductie van classificatietabellen met coveringgebaseerde ruwverzamelingen kritisch te bestuderen en de verbanden tussen verschillende definities en aanpakken in kaart te brengen.
Via dit onderwerp kan de student een rechtstreekse bijdrage leveren tot het actueel wetenschappelijk onderzoek in de wiskunde

---

Titel: Explainable methods for irony detection from tweets
Promotor: prof. Chris Cornelis (chris.cornelis@ugent.be); Begeleider: Olha Kaminska (olha.kaminska@ugent.be)

Korte beschrijving:
The branch of data science that works with text and learns how to extract the knowledge from it is called Natural Language Processing (NLP). Over the past decades, NLP has made significant headway in the field of sentiment analysis for social media, for instance in the identification of cyberbullying, or in the detection of different emotions expressed by tweets. A driving force in this evolution has been the use of deep learning approaches in the representation and classification of textual data. Deep learning can solve the aforementioned problems with remarkable accuracy, but also has an important drawback: as a black-box model, it does not provide any insight into how it came to a particular conclusion.Explainable AI (XAI) is an emerging field in machine learning that aims to address how AI systems make decisions. It refers to AI methods and techniques that produce human comprehensible solutions. The latter can either be achieved by adding an additional layer of interpretability to black box methods, or by using intrinsically interpretable machine learning methods.
In this thesis, we focus on irony detection from tweets, a particularly challenging task, sometimes even for humans. For example, the sentence "Great. Another rainy day. How wonderful." could easily be mistaken by a machine to express positive sentiment. Ideally, we would like to obtain a system that accurately predicts whether a given tweet is ironic, and at the same time provides the motivation for its decisions (for example, "rainy days" are typically not "great" and "wonderful").

We focus on the the solution of SemEval 2018 Task 3, "Irony detection in English tweets". The goal will be to build a binary classification system, predicting whether a proposed tweet in English is ironic. By contrast to the original SemEval competition, the quality of the solution will be evaluated not only based on its accuracy, but also on its ability to explain predictions. Therefore, a suitable trade-off between both characteristics needs to be sought.

The preferred language of programming for this task is Python.

Titel: Technieken voor experimentele algoritmiek
Promotor: prof. Veerle Fack (Veerle.Fack@ugent.be)

Korte beschrijving:
Het analyseren van algoritmen behelst het voorspellen hoe goed een algoritme zal presteren in een gegeven situatie met gegeven voorwaarden en veronderstellingen.
Het ontwerpen van algoritmen behelst het bouwen van betere algoritmen, zoals snellere algoritmen of algoritmen die een goede (bij voorkeur optimale) oplossing voor een probleem bekomen.
Het gebied van de experimentele algoritmiek combineert het theoretische domein van algoritmiek met de praktijk.

Traditionele technieken voor de analyse van algoritmen volgen een theoretisch model (RAM-model met kost 1 per bewerking) voor het bestuderen van het gedrag van algoritmen. Voor meer accurate inschattingen van het gedrag van een algoritme zijn meer verfijnde modellen nodig, die ook ondersteund worden door experimenten op actuele computersystemen. De kern van dergelijk 'algorithm engineering' is een cyclus van ontwerp, analyse, implementatie en experimenten.
Deze masterscriptie heeft tot doel de student vertrouwd te maken met deze cyclus van algoritme-ontwerp, via een reeks van case studies, gaande van enkele eenvoudige algoritmische problemen (zoals zoeken in gesorteerde lijsten en sorteren) tot meer realistische problemen (zoals routeplanning).

Titel: Algoritmen voor distributieproblemen
Promotor: prof. Veerle Fack (Veerle.Fack@ugent.be)

Korte beschrijving:
In deze masterproef bestuderen we enkele optimalisatieproblemen waarmee distributiebedrijven te maken krijgen.
In het Bin Packing Problem (BPP) beschouwen we het inpakken van een reeks objecten met verscheidene gewichten in een reeks containers met inhoud V, op zodanige manier dat zo weinig mogelijk containers gebruikt worden.
In het Vehicle Routing Problem (VRP) beschouwen we een pakjesdienst, die dagelijks goederen moet afleveren bij veel verschillende klanten. Hiervoor is een vloot van voertuigen beschikbaar, die opereert vanuit een centraal distributiecentrum. Het doel is een route voor elk voertuig te ontwerpen (vergelijkbaar met de route uit het handelreizigersprobleem), zodanig dat alle klanten bediend worden door precies een voertuig en dat de totale reiskost van de voertuigen minimaal is.
In het Facility Location Problem (FLP) beschouwen we een distributiebedrijf, dat goederen in bulk opslaat in meerdere opslagruimtes, om ze van daaruit te verdelen naar meerdere klanten. Het doel is om te bepalen welke opslagruimtes geschikt zijn voor welke goederen, zodanig dat de kost om de klanten te bedienen minimaal is. De moeilijkheid van dit probleem komt van het feit dat elke opslagruimte een eigen kost en een eigen opslagcapaciteit heeft .Het is de bedoeling om benaderingsalgoritmen (gebaseerd op metaheuristieken zoals local search, tabu search, genetische algoritmen, e.d.) uit te werken voor dergelijke distributieproblemen.

Titel: Algoritmen voor het Steiner Tree Problem
Promotor: prof. Veerle Fack (Veerle.Fack@ugent.be)

Korte beschrijving:
In een brede betekenis bestaat het doel van een Steiner Tree Problem erin om de goedkoopste manier te bepalen om een set van objecten te verbinden.
In de meest voorkomende varianten zijn deze objecten ofwel punten in een metrische ruimte ofwel een deelverzameling van de toppen van een netwerk/graaf, en het doel is het bepalen van een boom die ze allemaal verbindt.
Er zijn talloze toepassingen van deze problemen, zoals netwerkoptimalisatie, reconstructie van phylogenetische bomen, ontwerp van computercircuits, multicast-routering in communicatienetwerken.
Dit probleem was ook het onderwerp van een DIMACS Implementation Challenge in 2014, waarbij o.m.
een collectie benchmark data opgesteld werd, naast het uitwerken van gevorderde algoritmen voor het aanpakken van dit probleem.
Met deze Implementation Challenge als startpunt, is het de bedoeling van deze masterproef om bestaande algoritmen voor het probleem te bestuderen en te analyseren, evenals eigen benaderingen uit te werken en te toetsen tegenover de bestaande methodes

Similariteitsrelaties voor multi-instance data

Promotor: Chris Cornelis (chris.cornelis@ugent.be)

Situering

In een traditionele classificatie-dataset wordt elke observatie beschreven aan de hand van zijn attribuutwaarden en een bijhorend klasselabel. Tabel 1 bevat enkele observaties uit de bekende iris-dataset. De observaties i_1, i_2 en i_3 worden beschreven met vier attributen (sepalLength, sepalWidth, petalLength, petalWidth) die eigenschappen van een irisplant voorstellen. Er zijn drie mogelijke klasselabels (soorten iris): setosa, versicolor en virginica.

ID	sepalLength	sepalWidth	petalLength	petalWidth	Class
i_1	5.1	3.5	1.4	0.2	setosa
i_2	6.0	2.2	4.0	1.0	versicolor
i_3	6.2	3.4	5.4	2.3	virginica

Tabel 1: Enkele observaties uit de dataset Iris.

In multi-instance data komt elke observatie overeen met een groep elementen. Zo'n groep wordt een *bag* genoemd, en het aantal elementen kan verschillen van bag tot bag. Elk element kan worden voorgesteld met een attribuutvector, maar heeft geen afzonderlijk klasselabel. Het label wordt toegekend aan de gehele bag. Tabel 2 toont twee observaties uit de multi-instance dataset Musk, waarbij elke bag diverse conformaties van eenzelfde molecule bevat. De eerste bag behoort tot de positieve klasse (muskusgeur), de tweede tot de negatieve klasse (geen muskusgeur).

Bag ID	$\langle f_1, f_2, \ldots, f_{m-1}, f_m \rangle$	Class
MUSK-jf59	$\langle 52, -110, \ldots, -60, -29 \rangle$ $\langle 49, -98, \ldots, -13, -12 \rangle$ $\langle 23, -113, \ldots, -9, 90 \rangle$ $\langle 47, -110, \ldots, -5, -8 \rangle$ $\langle 9, -114, \ldots, -28, 112 \rangle$	Positive
NON-MUSK-334	$\langle 7, -197, \ldots, 34, 55 \rangle$ $\langle 25, -198, \ldots, 20, -8 \rangle$	Negative

Tabel 2: Enkele observaties uit de multi-instance dataset Musk.

Probleemstelling

Een cruciale component van veel classificatie-algoritmen is het gebruik van afstand of similariteit. Nearest neighbor methoden bijvoorbeeld voorspellen het klasselabel van een nieuwe observatie door diens dichtste buren in de dataset op te sporen, en het meest voorkomende klasselabel onder de buren te kiezen. In de Iris-dataset kan dit zoekproces gebeuren door een metriek (bvb. Euclidische afstand) te evalueren op de attribuutvectoren van observaties.
Multi-instance data zijn duidelijker complexer dan klassieke data. Hoe bepalen we bijvoorbeeld de afstand of mate van gelijkenis tussen twee bags? Er bestaan in de wetenschappelijke literatuur diverse voorstellen om dit probleem op te lossen. In deze masterproef zullen we de mate van gelijkenis tussen multi-instance bags voorstellen aan de hand van *vaagrelaties*, d.w.z. afbeeldingen van $\mathcal{B}$ naar $[0, 1]$ waarbij $\mathcal{B}$ de klasse der bags uit het gestelde classificatieprobleem voorstelt.

Doelstelling

Het doel van deze masterproef omvat:

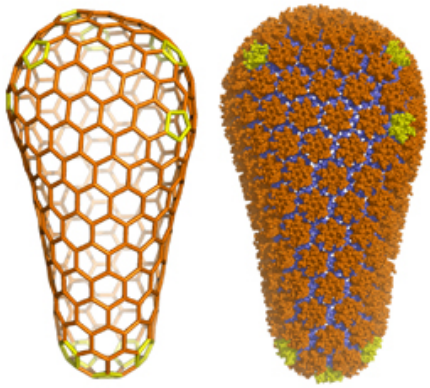
1. Een grondige en kritische literatuurstudie van de bestaande voorstellen, gezien als vaagrelaties en ingedeeld volgens hun constructiemethode en de eigenschappen (bvb. diverse vormen van transitiviteit) die ze vervullen.

2. Afhankelijk van je interesse, ofwel (a) een experimentele studie waarbij je de belangrijkste methoden vergelijkt in een classificatie-experiment, ofwel (b) een theoretische studie waarbij je zelf één of meerdere nieuwe vaagrelaties voorstelt en onderzoekt, of (c) een combinatie van beide.

Clusters in Fullerenen en HIV virussen

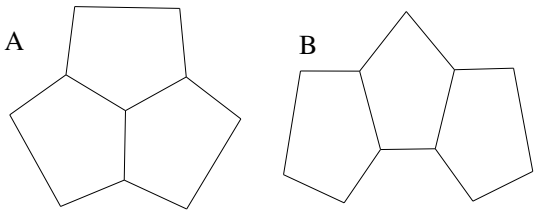
Promotor: Gunnar Brinkmann

Fullerenen en HIV virussen hebben vanuit een wiskundig oogpunt een gelijkaardige structuur: het zijn 3-reguliere vlakke grafen met alleen maar 5- en 6-hoeken.



Een pentagoncluster in een Fullereen of HIV virus is een maximale groep van vijfhoeken die samenhangen. Er is maar één mogelijke cluster met één of twee vijfhoeken, er zijn twee clusters met drie vijfhoeken, vier clusters met vier vijfhoeken, etc.

In een Fullereen of HIV virus moet de som van de groottes van de clusters altijd 12 zijn, maar als de som van de groottes 12 is, betekent dat niet dat er ook een Fullereen of HIV virus met deze clusters bestaat – zelfs dan niet als je alleen maar een theoretisch mogelijke structuur zoekt en geen rekening houdt met scheikundige of biologische beperkingen. Als één van de clusters heel groot is, is het gemakkelijk gevallen te vinden waar een Fullereen met die clusters niet kan bestaan. Voor kleine clusters ligt het vaak niet voor de hand welke combinaties wel dan niet mogelijk zijn. Is er bv. een Fullereen of HIV virus die drie clusters van het type A en één cluster van het Type B uit de volgende afbeelding bevat?



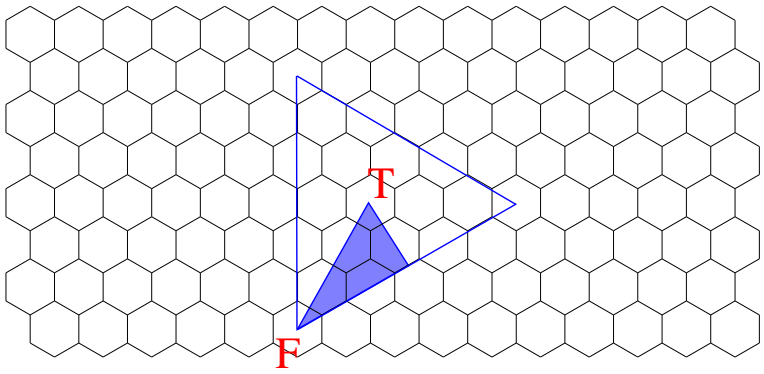
In het kader van deze thesis zal door een combinatie van computationele en theoretische technieken onder andere bepaald worden welke combinaties van pentagonclusters in Fullerenen kunnen voorkomen.

Meer uitleg kan natuurlijk aan de promotor gevraagd worden. Het is zinvol het vak *algoritmische grafentheorie* gevolgd te hebben.

Goldbergoperaties voor clusters

Promotor: Gunnar Brinkmann

Michael Goldberg heeft in 1934 een operatie gedefinieerd om 3-reguliere vlakke grafen met icosahedrale symmetrie en alleen maar 5-hoeken en 6-hoeken te construeren. Later werd een constructie Goldberg-operatie genoemd en heel populair omdat sommige virussen en vooral de in 1985 ontdekte Fullerenen een dergelijke structuur hebben. Inderdaad werd deze Goldberg-operatie door de biologen Caspar en Klug geïntroduceerd, wij noemen die dus beter Caspar-Klug-operaties. Zowel door middel van (de echte) Goldbergoperaties en de Caspar-Klug-operaties kan je van een Fullereen een groter Fullereen maken met dezelfde symmetriegroep. Goldberg gebruikt daarbij een driehoek gebied uit de hexagonale tralie. Deze Goldberg driehoek voor de parameters $(3, 3)$ zie je in het blauw in de afbeelding. Als je meerdere kopieën van deze driehoek (en de gespiegelde driehoek) uitknipt en op de juiste manier in de vlakken van een Fullereen plakt, dan kan je een Fullereen met 27 keer meer toppen krijgen.



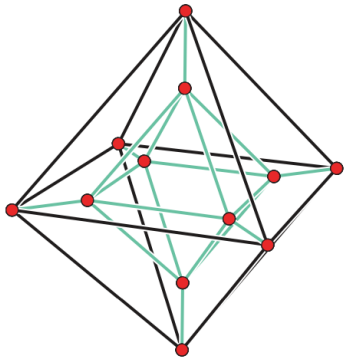
In het resultaat zullen de 12 vijfhoeken altijd geïsoleerd zijn van elkaar. Het zou mooi zijn ook een (op de Goldbergoperaties opbouwende) operatie te hebben die clusters van driehoeken behoudt – waar er dus als er bv. in het oorspronkelijke Fullereen drie vijfhoeken rond een gemeenschappelijke top zitten dat ook in het grotere Fullereen zo is. Dat kan voor sommige (maar niet alle) Goldberg operaties inderdaad door achteraf delen van het Fullereen te vervangen. In het kader van deze thesis worden de details van deze nieuwe operaties uitgewerkt en wordt bewezen dat de nieuwe operaties de gewenste eigenschappen hebben.

Meer uitleg kan natuurlijk aan de promotor gevraagd worden. Het is zinvol het vak *algoritmische grafentheorie* gevolgd te hebben.

Polyhedrale grafen met een niet-hamiltoniaans prisma

Promotoren: Jan Goedgebeur en Carol T. Zamfirescu

Dit project gaat over de hamiltoniaanse eigenschappen van het cartesisch product van het 1-skelet van een 3-polytoop (equivalent, een planaire 3-samenhangende graaf) met K_2 (de complete graaf met 2 toppen), d.w.z. zijn *prisma*. Figuur 1 toont de graaf van het prisma van een octahedron. Wanneer we hier spreken over een d-polytoop (d.w.z. een d-dimensionale polytoop), hebben we het altijd over zijn 1-skelet (dus een graaf). Een d-polytoop heet *simpel* of *eenvoudig* als elk van zijn toppen incident is met exact d bogen. Door een stelling van Balinski weten we dat een d-polytoop d-samenhangend moet zijn, dus dat alle toppen minimaal graad d hebben.



Figuur 1: Het prisma van een octahedron.

Rosenfeld en Barnette hebben bewezen dat de 4-kleuren-stelling impliceert dat eenvoudige 3-polytopen hamiltoniaanse prisma's hebben. Fleischner heeft een bewijs gevonden om het gebruik van de 4-kleuren-stelling te vermijden. Paulraja versterkte dit door aan te tonen dat dezelfde conclusie geldt, zelfs als de planariteit niet wordt aangenomen, d.w.z. dat het prisma van een 3-samenhangende kubische graaf hamiltoniaans is. Voor een recente samenvatting van verdere resultaten op het gebied van prisma-hamiltoniciteit, verwijzen we naar [1].

In 1973 formuleerden Rosenfeld en Barnette het vermoeden dat elk 3-polytoop een hamiltoniaans prisma heeft. Špacapan [1] heeft dit vermoeden recentelijk weerlegd. Maar de vraag of alle simpele 4-polytopen hamiltoniaans zijn, blijft open. Deze vraag is de titel van een artikel van Rosenfeld uit 1983. Daarin construeert hij uitgebreide families van niet-hamiltoniaanse 4-reguliere 4-samenhangende grafen waarvan geen enkele het 1-skelet van een 4-polytoop is. Al in 1970 formuleerde Barnette het vermoeden dat het antwoord op deze vraag positief is. Aan de andere kant denkt Mohar dat er voor elke gehele $d \geq 3$ een simpele d-polytoop bestaat die niet hamiltoniaans is. Voor $d = 3$ werd dit bewezen door Tutte in 1946. We merken op dat als men de simpelheid laat vallen, niet-hamiltoniaanse d-polytopen vrij gemakkelijk te beschrijven zijn (bijvoorbeeld met behulp van Kleetopes).

We zien dat in dit probleem, zoals in veel vragen met betrekking tot de hamiltoniaanse eigenschappen van grafen, de graden van toppen een cruciale rol spelen. In een recente samenwerking met Daiki en Maezawa hebben we daarom de oplossing van Špacapan [1] aangepast om enkele natuurlijke vragen aan te pakken, in het bijzonder: hoe laag kunnen we gaan met de maximale graad? Hoeveel toppen van graad 4 kunnen we forceren? Kunnen we de conclusie versterken dat de familie van de 3-polytopen niet prisma-hamiltoniaans is?

De bedoeling van deze thesis is om dit soort vragen vanuit zowel een theoretisch als een computationeel oogpunt aan te pakken. In het bijzonder is het de bedoeling om algoritmes te ontwerpen en te implementeren om dit soort grafen te construeren en hun hamiltoniaanse eigenschappen te bestuderen.

Voor meer uitleg, contacteer: Jan.Goedgebeur@UGent.be en Carol.Zamfirescu@UGent.be

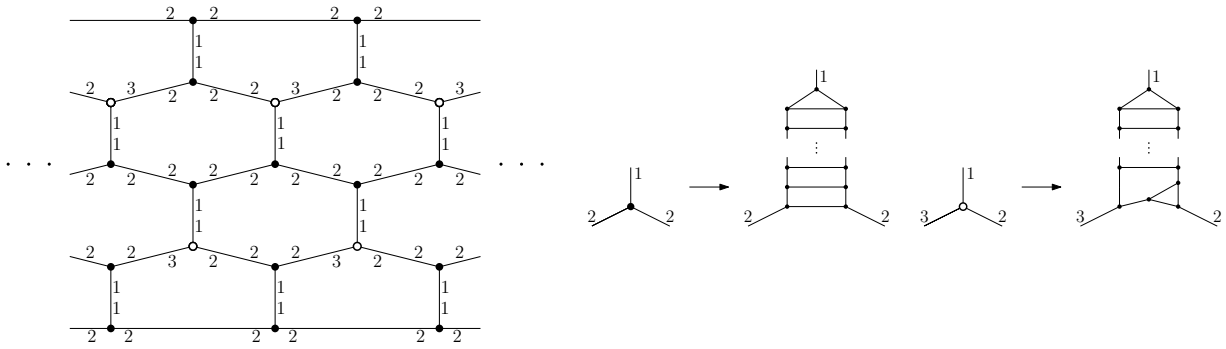
Referenties

[1] S. Špacapan. A counterexample to prism-hamiltonicity of 3-connected planar graphs. arXiv:1906.06683.

Cykellengtes in 3-samenhangende grafen

Promotor: Carol T. Zamfirescu

In dit onderwerp bestuderen we, voor een gegeven graaf G, de verzameling $\mathcal{C}(G)$ van cykellengtes in G, dus het *cykelspectrum* van G. Merker [2] bewees onlangs dat voor elk natuurlijk getal k elke 3-samenhangende 3-reguliere planaire graaf G met *omtrek* – dus de lengte van een langste cykel – minstens k voldoet aan $\mathcal{C}(G) \cap [k, 2k+9] \neq \emptyset$. Alhoewel het bestuderen van cykels in planaire grafen een lange geschiedenis heeft was dit resultaat van Merker interessant en nieuw. Merker poneerde in zijn artikel het vermoeden dat voor elk natuurlijk getal $k \geq 2$, elke 3-samenhangende 3-reguliere planaire graaf G met omtrek minstens k voldoet aan $\mathcal{C}(G) \cap [k, 2k+2] \neq \emptyset$. Het is niet zo moeilijk – en misschien een goede oefeningen om op te warmen – te zien dat dit geldt voor alle $k \in \{2, 3, 4, 5\}$.



De figuren tonen de constructie van 3-samenhangende 3-reguliere planaire grafen met een groot gat in hun cykelspectrum.

Maar recent werd er bewezen dat er voor elk even getal $k \geq 6$ oneindig veel tegenvoorbeelden voor dit vermoeden bestaan [3]. Cui en Lo vulden dit aan door de intervallen voor elke k volledig te beschrijven [1]. De bedoeling van deze thesis is om antwoorden op een aantal vragen in de context van deze resultaten te geven; bijvoorbeeld kunnen al deze vragen gesteld worden voor bepaalde deelverzamelingen van de familie van alle 3-samenhangende planaire grafen, of voor andere oppervlakken dan de bol (equivalent: planaire grafen) – het bestuderen van de situatie op de torus zou al interessante resultaten kunnen opleveren.

Voor meer uitleg contacteer: Carol.Zamfirescu@UGent.be.

Referenties

[1] Q. Cui en O.-H. S. Lo. Tight gaps in the cycle spectrum of 3-connected planar graphs. arXiv:2009.02503 [math.CO].

[2] M. Merker. Gaps in the cycle spectrum of 3-connected cubic planar graphs. *J. Combin. Theory, Ser. B* **146** (2021) 68–75.

[3] C. T. Zamfirescu. Counterexamples to a conjecture of Merker on 3-connected cubic planar graphs with a large cycle spectrum gap. arXiv:2009.00423 [math.CO].