Liduine Kampschöer

Speltheoretisch optimaal voorspellen voor Bregman divergenties

Bachelorscriptie

Scriptiebegeleider: Dr. T. van Erven

Datum bachelorexamen: 3 juli 2015

Mathematisch Instituut, Universiteit Leiden

Inhoudsopgave

1 Inleiding

Machine learning is een deelgebied van de statistiek, waarbij voorspellen centraal staat. In de huidige tijd is het maken van precieze voorspellingen op veel verschillende vlakken erg belangrijk. In deze bachelorscriptie zal worden geconcentreerd op één specifiek geval, waarbij de kwaliteit van voorspellingen van groot belang is.

Électricité de France, het grootste elektriciteitsbedrijf van Frankrijk, produceert elke dag een hoeveelheid elektriciteit, gebaseerd op een voorspelling, die vóór middernacht is gemaakt, van hoeveel elektriciteit er de volgende dag nodig zal zijn in het land. Wanneer er te veel elektriciteit geproduceerd wordt, gaat elektriciteit verloren, omdat dit lastig opgeslagen kan worden. Wanneer er te weinig elektriciteit geproduceerd wordt, moet Électricité de France dit op het laatste moment duur inkopen bij een buitenlandse partner of binnenlandse concurrent. Voor efficiëntie is een nauwkeurige voorspelling dus noodzakelijk.

Het blijkt dat over het algemeen originele voorspellingen inconsistent zijn. In deze scriptie bestuderen we een voorspelling voor het totale elektriciteitsverbruik en voorspellingen voor het elektriciteitsverbruik van verschillende deelgroepen, waarbij de deelgroepen bij elkaar opgeteld het totaal vormen. Inconsistentie wil nu zeggen dat de voorspelling van het totaal niet overeenkomt met de som van de voorspellingen van alle deelgroepen. We zullen gebruikmaken van een methode, gebaseerd op speltheorie, om deze voorspellingen consistent te maken, wat dus betekent dat het voorspelde totaal wel overeenkomt met de som van de voorspelde deelgroepen.

Het consistent maken van schattingen wordt al jaren uitgebreid bestudeerd vanwege de vele toepassingen. Al in 1942 stelde Stone dat het essentieel was dat het totaalbedrag en de verschillende posten bij elkaar opgeteld overeenkwamen in de nationale begroting van Engeland [4]. Ook Hyndman hield zich met consistentie bezig. Toegepast op data over het Australische toerisme, moest volgens hem het totaal aantal toeristen wel overeenkomen met de som van het aantal toeristen per gebied [3].

Wanneer we met behulp van de originele, inconsistente voorspellingen nieuwe, consistente voorspellingen maken, willen we dat hiermee nooit de kwaliteit verslechtert. In eerste instantie meten we de voorspelkwaliteit met de kwadratische Euclidische afstand. In stelling 1 tonen we aan dat hiervoor inderdaad zo'n voorspelling blijkt te bestaan, die voldoet aan dit criterium. Dit is gebaseerd op een bestaand resultaat [5]. We willen echter ook voorspellingen in de vorm van kansverdelingen kunnen doen, aangezien hier in de praktijk vraag naar is. Daarvoor moeten we de kwaliteit meten met een maat die geschikt is voor kansverdelingen. De standaard maat hiervoor is de Kullback-Leibler divergentie, die we dan ook zullen gebruiken.

Voor we de Kullback-Leibler divergentie kunnen toepassen op ons probleem, moeten we eerst de kwadratische Euclidische afstand generaliseren naar de algemene groep van kwaliteitsmaten, de Bregman divergenties, waarvan de kwadratische afstand de bekendste is. We meten de kwaliteit van voorspellingen

met behulp van deze Bregman divergenties. Wat we eerder gesteld hebben over een bestaande voorspelling voor de kwadratische afstand waarbij de kwaliteit nooit verslechtert, generaliseren we in stelling 2 naar Bregman divergenties in het algemeen. Dit is het eerste nieuwe resultaat.

Wanneer we de generalisatie hebben geformuleerd, kunnen we hiervandaan specialiseren naar één specifieke Bregman divergentie, in ons geval naar de Kullback-Leibler divergentie. Dit resultaat zal geformuleerd worden in stelling 3.

In het volgende hoofdstuk zullen definities, zoals de consistentie en de kwaliteit van voorspellingen, geïntroduceerd worden. Op basis hiervan kan stelling 1 gepresenteerd worden, in de laatste paragraaf van dit hoofdstuk. In hoofdstuk 3 komen Bregman divergenties aan de orde. De definitie van deze verzameling van kwaliteitsmaten en de generalisatie, stelling 2, kunnen we hierin terugvinden. In het vierde hoofdstuk bekijken we deze stelling in geval van de Kullback-Leibler divergentie, de divergentie die het verschil tussen twee kansverdelingen meet. Het laatste hoofdstuk bestaat uit een samenvatting en een beschouwing van mogelijk toekomstig onderzoek.

2 Een speltheoretisch optimale methode

Een speltheoretische methode voor dit probleem is praktisch, aangezien er dan helemaal geen aannames hoeven worden gedaan. De eigenschappen van een kansverdeling, zoals de variantie en covariantie, doen er niet toe; in ieder geval kan deze methode worden toegepast. Hoe werkt deze speltheoretisch optimale methode?

2.1 Voorspellingen

In ons geval van elektriciteitsverbruik van Électricité de France willen we twee verschillende dingen voorspellen. Allereerst zijn we geïnteresseerd in het *elektriciteitsverbruik per deelgroep*. Een deelgroep kan gezien worden als elke soort groep klanten die elektriciteit consumeren, zoals bijvoorbeeld een aantal bedrijven met hetzelfde contract. Voor de eenvoud en zodat we het een naam kunnen geven, gaan wij er in de rest van deze scriptie vanuit dat een deelgroep op een bepaalde regio in Frankrijk slaat. Stel dat Frankrijk uit K regio's bestaat, willen wij het elektriciteitsverbruik in deze K regio's voorspellen. Daarnaast zijn we geïnteresseerd in het *totale elektriciteitsverbruik* van deze regio's. Hoeveel elektriciteit zal er totaal morgen in Frankrijk nodig zijn?

2.2 Consistentie

Het totale daadwerkelijke elektriciteitsverbruik, Y_{tot}, en het elektriciteitsverbruik van de regio's, $Y_1, Y_2, \ldots, Y_K$, zijn consistent. Dit wil zeggen dat het totale verbruik inderdaad overeenkomt met de som van het gebruik van alle regio's. Er geldt:

$$Y_{tot} = Y_1 + Y_2 + \ldots + Y_K.$$

Over het algemeen blijkt echter dat bij het voorspellen van elektriciteitsverbruik, het totale voorspelde verbruik, $\hat{Y}_{tot}$, níet overeenkomt met de som van

het voorspelde gebruik van alle regio's, $\hat{Y}_1 + \hat{Y}_2 + \ldots + \hat{Y}_K$. Het totale verbruik kan namelijk beter voorspeld worden dan enkel de voorspellingen per regio bij elkaar op te tellen. Hier zijn meerdere verklaringen voor.

Sowieso zijn er bepaalde factoren die we alleen kunnen meenemen bij het voorspellen van het elektriciteitsverbruik van heel Frankrijk. Stel even, in tegenstelling tot in de rest van deze scriptie, dat K het aantal verschillende groepen klanten met een verschillend contract is. In de zomervakantie wordt over het algemeen minder elektriciteit in het hele land gebruikt; dit beïnvloedt de voorspelling van het totale verbruik. Per gebied in Frankrijk is bekend wanneer daar zomervakantie is en dus wanneer daar waarschijnlijk minder elektriciteit gebruikt zal worden, door bijvoorbeeld gesloten scholen en bedrijven of mensen die op vakantie zijn. Dat gaat echter niet op voor alle uitzonderlijke klantengroepen met een verschillend contract. De klanten binnen zo'n klantengroep zijn verspreid over het hele land. Er is dus geen periode gebaseerd op vakantiedata, waarover we zouden kunnen zeggen dat er minder elektriciteit verbruikt zal worden binnen deze klantengroep. Oftewel, er is geen dergelijke extra informatie die we zouden kunnen gebruiken bij de afzonderlijke voorspellingen, $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_K$.

Ook is er een statistische verklaring. Tegenwoordig worden allerlei moderne methodes toegepast, waarbij onzuivere schatters gebruikt worden, omdat deze erg nauwkeurige schattingen produceren. Deze schatters verminderen de variantie in hun voorspellingen op een manier die ervoor zorgt dat ze gemiddeld een klein beetje naast de werkelijke waarde zitten, oftewel dat ze een klein beetje onzuiver zijn. Voor de grootste nauwkeurigheid zal dus inconsistentie moeten gelden: $\mathbb{E}[\hat{Y}_k] \neq \mathbb{E}[Y_k]$.

Als laatste zouden we ons kunnen voorstellen dat verschillende effecten bij aparte regio's, zoals een beetje meer dan gemiddeld verbruikte elektriciteit bij de één en een beetje minder dan gemiddeld verbruikte elektriciteit bij de ander, uitmiddelen bij het optellen ervan. Bij de voorspellingen kunnen we hier echter geen rekening meehouden. Dit zou als derde verklaring van de inconsistentie van $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_K, \hat{Y}_{tot})$ gezien kunnen worden.

Samenvattend hebben we in het geval van voorspellingen van elektriciteitsverbruik dus te maken met inconsistentie:

$$\hat{Y}_{tot} \neq \hat{Y}_1 + \hat{Y}_2 + \ldots + \hat{Y}_K.$$

Zoals men intuïtief al zou denken, brengt inconsistentie problemen met zich mee. Hoe moet de totaal voorspelde elektriciteit verdeeld worden, als deze niet overeenkomt met de som van de voorspelde elektriciteit van alle regio's? Inconsistentie in voorspellingen wordt dan ook voor operationele redenen door managers vaak niet aanvaard. Hier moet dus een oplossing voor worden gezocht.

2.3 Van inconsistentie naar consistentie

Voor de hand ligt misschien om gewoon de voorspellingen $\hat{Y}_1, \ldots, \hat{Y}_K$ bij elkaar op te tellen en zo op een totale voorspelling te komen: $\hat{Y}_1 + \ldots + \hat{Y}_K = \hat{Y}_{tot}$. Dit

wordt de zogenaamde *bottom-up methode* genoemd. De voorspellingen zijn dan wel consistent, maar er wordt überhaupt geen gebruik meer gemaakt van onze kennis over de totale voorspelling. Extra kennis om de voorspellingen preciezer te maken zou uiteraard niet weggegooid mogen worden. Immers, hoe beter de voorspellingen zijn, hoe meer deze waard zijn voor Électricité de France.

Machine learners produceren inconsistente voorspellingen, terwijl managers consistente voorspellingen eisen. Aangezien we onze inconsistente, vaak veel betere voorspellingen wel willen gebruiken, zetten we deze om naar nieuwe, consistente en daarmee bruikbare, voorspellingen. We beelden $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_K, \hat{Y}_{tot})$ af op de nieuwe voorspellingen $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_K, \tilde{Y}_{tot})$ waarvoor wel geldt:

$$\tilde{Y}_{tot} = \tilde{Y}_1 + \ldots + \tilde{Y}_K.$$

2.4 De kwaliteit van voorspellingen

Het verlies ten opzichte van het daadwerkelijke verbruik $\mathbf{Y}$, en daarmee de kwaliteit, van voorspellingen $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_K, \hat{Y}_{tot})$ kunnen we bijvoorbeeld meten met de kwadratische Euclidische afstand. Dit is de meest voorkomende afstandsmaat in de statistiek en wordt dan ook gebruikt bij Électricité de France. Er geldt:

$$\ell(\mathbf{Y}, \hat{\mathbf{Y}}) = \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2 = \sum_{k=1}^{K} (\hat{Y}_k - Y_k)^2 + (\hat{Y}_{tot} - Y_{tot})^2. \tag{1}$$

Hoe kleiner de afstand tussen $\hat{\mathbf{Y}}$ en $\mathbf{Y}$ is, des te kleiner het verschil hiertussen en daarmee des te hoger de kwaliteit van de voorspellingen.

Nu we een afbeelding maken van $\hat{\mathbf{Y}}$ naar $\tilde{\mathbf{Y}}$ willen we dat hierbij de voorspelkwaliteit nooit verslechtert. De consistente voorspellingen $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_K, \tilde{Y}_{tot})$ moeten voor elk verbruik $\mathbf{Y}$ minstens even goed zijn als de voorspellingen $\hat{\mathbf{Y}}$, hierboven beschreven. Voor de kwaliteit van de consistente voorspellingen $\tilde{\mathbf{Y}}$ geldt analoog aan (1): $\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) = \| \tilde{\mathbf{Y}} - \mathbf{Y} \|^2 = \sum_{k=1}^{K} (\tilde{Y}_k - Y_k)^2 + (\tilde{Y}_{tot} - Y_{tot})^2$. Het verschil tussen de voorspelkwaliteit van de nieuwe, consistente voorspellingen en de originele, inconsistente voorspellingen is dus $\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})$. Aangezien we willen dat $\ell(\mathbf{Y}, \tilde{\mathbf{Y}})$ niet groter is, willen we dat $\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})$ niet positief is. Omdat het onbekend is wat $\mathbf{Y}$ zal zijn en we willen dat voor elke $\mathbf{Y}$ geldt dat het bovenstaande verschil hoogstens nul is, moeten we rekening houden met het ergste geval. Dit is de kleinste bovengrens, oftewel het supremum van het verschil:

$$\sup_{\mathbf{Y} \in S} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})),$$

waarbij $S = \{\mathbf{X} \in \mathbb{R}^{K+1} | \sum_{k=1}^{K} X_k = X_{K+1}\}$, de verzameling van consistente vectoren. Als het voor het ergste geval zelfs op gaat dat $\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}}) \leq 0$, geldt dus voor alle gevallen, voor elk verbruik $\mathbf{Y}$, dat voorspellingen $\tilde{\mathbf{Y}}$ niet slechter zijn dan voorspellingen $\hat{\mathbf{Y}}$.

$\tilde{\mathbf{Y}}$ moet zo gekozen worden dat zelfs het grootste verschil in $\sup_{\mathbf{Y} \in S} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}}))$ nog negatief is. Het maximale verschil willen we dus minimaliseren.

Dit is een zogenaamd minimax optimalisatie probleem. De speltheoretisch optimale manier hiervoor is om de voorspellingen $\tilde{\mathbf{Y}}$ zo te kiezen dat deze het minimum bereiken in

$$V = \min_{\tilde{\mathbf{Y}} \in S} \max_{\mathbf{Y} \in S} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})). \tag{2}$$

2.5 Speltheorie

We gebruiken een speltheoretisch optimale methode waarbij een verzameling van inconsistente voorspellingen gebruikt wordt als input en als output een verzameling van consistente voorspellingen wordt geproduceerd, die ten minste even goed zijn.

Voor de volgende stelling verwijzen we naar theorie 1 van [5]:

Stelling 1. *Er bestaat de unieke projectie* $\tilde{\mathbf{Y}}_{\mathbf{proj}} = \arg\min_{\tilde{\mathbf{Y}} \in S} \parallel \tilde{\mathbf{Y}} - \hat{\mathbf{Y}} \parallel^2$*, waarbij*

$$V = - \parallel \tilde{\mathbf{Y}}_{\mathbf{proj}} - \hat{\mathbf{Y}} \parallel^2 \leq 0,$$

en waar voor de speltheoretisch optimale voorspellingen geldt $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_{\mathbf{proj}}$.

In deze uitdrukking impliceert $V \leq 0$ dat deze nieuwe voorspellingen minstens even goed zijn als de originele voorspellingen:

$$\forall \mathbf{Y} \in S : \ell(\mathbf{Y}, \tilde{\mathbf{Y}}_{\mathbf{proj}}) \leq \ell(\mathbf{Y}, \hat{\mathbf{Y}}).$$

Deze stelling zal later in deze paragraaf worden bewezen.

Wij bekijken nu een spel tussen twee spelers, waarbij elke speler aan het eind van het spel een score krijgt, die hij of zij zo groot mogelijk wil hebben. Dit wordt ook wel de pay off genoemd. Bij het beschouwen van een *zero-sum game* geldt dat in iedere eindsituatie de waarde van speler 1 tegenovergesteld is aan de waarde van speler 2. Als we de pay offs voor speler 1 en 2 respectievelijk u_1 en u_2 noemen geldt dat de som hiervan gelijk is aan nul: $u_1 + u_2 = 0$. Dit verklaart de naam zero-sum game. Voor ons probleem hebben we te maken met de volgende situatie:

- Het betreft een zero-sum game;
- Beide spelers hebben één beurt;
- Mogelijke zetten voor spelers hangen niet af van wat de ander doet.

Speler 1 kiest een zet $a \in A$ en speler 2 kiest een zet $b \in B$, waarbij A en B respectievelijk de verzamelingen van alle mogelijke zetten voor speler 1 en 2 zijn. Omdat we een zero-sum game beschouwen, kunnen we beide pay offs uitdrukken met behulp van een enkele functie f. Dit geeft pay off $u_1 = f(a, b)$ en $u_2 = -f(a, b)$ voor respectievelijk speler 1 en 2. Het uitgangspunt is dat beide spelers hun pay off optimaliseren. Wanneer speler 1 eerst aan de beurt is geldt dat zijn of haar pay off gelijk is aan $\max_{a \in A} \min_{b \in B} f(a, b)$. Als speler 2 eerst aan de beurt is geldt dat de pay off van speler 1 gelijk is aan $\min_{b \in B} \max_{a \in A} f(a, b)$.

Voor de pay off van speler 2 geldt dus als speler 1 begint $-\max_{a\in A}\min_{b\in B} f(a,b)$ en als speler 2 begint $-\min_{b\in B}\max_{a\in A} f(a,b)$.

De speler die als tweede gaat, is altijd in het voordeel, aangezien deze nu weet welke zet de ander heeft gedaan. Wiskundig beschrijven we dit als volgt:

Lemma 1. *Voor iedere functie f geldt* $\sup_{a\in A}\inf_{b\in B} f(a,b) \leq \inf_{b\in B}\sup_{a\in A} f(a,b)$.

Bewijs. Voor willekeurige functie f geldt:

$$\forall a_0, b_0 : f(a_0, b_0) \leq \sup_{a\in A} f(a, b_0).$$

We nemen aan beide kanten het infimum over alle b:

$$\forall a_0 : \inf_{b\in B} f(a_0, b) \leq \inf_{b\in B}\sup_{a\in A} f(a,b).$$

Dan geldt ook:

$$\sup_{a\in A}\inf_{b\in B} f(a,b) \leq \inf_{b\in B}\sup_{a\in A} f(a,b).$$

□

Spelers kunnen volgens verschillende strategieën spelen. Een zadelpunt (a^*, b^*) is een punt waarbij beide spelers een theoretisch optimale strategie volgen.

Definitie 1. *Een* zadelpunt *van f is een punt (a^*, b^*) waarvoor geldt:*

$$\forall a : f(a, b^*) \leq f(a^*, b^*) \tag{3}$$
$$\forall b : f(a^*, b) \geq f(a^*, b^*). \tag{4}$$

Er geldt dat als één van de spelers volgens het zadelpunt speelt, de ander het ook niet erg vindt om volgens het zadelpunt te spelen.

Lemma 2. *Als er een zadelpunt (a^*, b^*) bestaat voor een functie f, geldt de minimax gelijkheid:*

$$M = \max_{a\in A}\min_{b\in B} f(a,b) = \min_{b\in B}\max_{a\in A} f(a,b), \tag{5}$$

en geldt bovendien dat deze gelijkheid ook gelijk is aan de waarde van f op het zadelpunt:

$$M = f(a^*, b^*). \tag{6}$$

Bewijs. Voor (a^*, b^*) een zadelpunt geldt:

$$f(a^*, b^*) = \max_{a\in A} f(a, b^*) \geq \min_{b\in B}\max_{a\in A} f(a,b),$$
$$f(a^*, b^*) = \min_{b\in B} f(a^*, b) \leq \max_{a\in A}\min_{b\in B} f(a,b),$$

dus geldt:

$$\min_{b\in B}\max_{a\in A} f(a,b) \leq f(a^*, b^*) \leq \max_{a\in A}\min_{b\in B} f(a,b) \leq \min_{b\in B}\max_{a\in A} f(a,b),$$

waarbij de laatste ongelijkheid geldt wegens lemma 1. Hieruit volgt dat alle $\leq$-tekens vervangen kunnen worden door gelijkheden en dit geeft ons (5) en (6). □

De basis voor het bewijs van stelling 1 is nu gelegd.

Bewijs stelling 1. Stel de minimax gelijkheid geldt voor de functie

$$f(\mathbf{Y}, \tilde{\mathbf{Y}}) = \| \tilde{\mathbf{Y}} - \mathbf{Y} \|^2 - \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2, \tag{7}$$

waarbij $\hat{\mathbf{Y}}$ vast, aangezien dit de oorspronkelijke voorspellingen zijn:

$$\min_{\tilde{\mathbf{Y}} \in S} \max_{\mathbf{Y} \in S} (\| \tilde{\mathbf{Y}} - \mathbf{Y} \|^2 - \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2) = \max_{\mathbf{Y} \in S} \min_{\tilde{\mathbf{Y}} \in S} (\| \tilde{\mathbf{Y}} - \mathbf{Y} \|^2 - \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2).$$

Dit kunnen we herschrijven:

$$\begin{aligned} \max_{\mathbf{Y} \in S} \min_{\tilde{\mathbf{Y}} \in S} (\| \tilde{\mathbf{Y}} - \mathbf{Y} \|^2 - \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2) &= \max_{\mathbf{Y} \in S} (- \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2) \\ &= - \min_{\mathbf{Y} \in S} (\| \hat{\mathbf{Y}} - \mathbf{Y} \|^2) = - \| \hat{\mathbf{Y}} - \tilde{\mathbf{Y}}_{\mathbf{proj}} \|^2, \end{aligned}$$

waarbij $\tilde{\mathbf{Y}}_{\mathbf{proj}}$ zoals in stelling 1.
Als dus nu geldt dat

$$V = \min_{\tilde{\mathbf{Y}} \in S} \max_{\mathbf{Y} \in S} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})) = \max_{\mathbf{Y} \in S} \min_{\tilde{\mathbf{Y}} \in S} ((\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}}))$$

zijn we klaar, aangezien we hebben aangetoond dat

$$\max_{\mathbf{Y} \in S} \min_{\tilde{\mathbf{Y}} \in S} ((\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})) = - \| \hat{\mathbf{Y}} - \tilde{\mathbf{Y}}_{\mathbf{proj}} \|^2 \leq 0.$$

We gaan de stappen van de bovenstaande herschrijving vertalen naar zetten van spelers. Speler 1 kiest om f te maximaliseren $\mathbf{Y} = \tilde{\mathbf{Y}}_{\mathbf{proj}}$. Speler 2 kiest om f te minimaliseren $\tilde{\mathbf{Y}} = \mathbf{Y}$. Uit de zet van speler 1, volgt de keuze voor de zet van speler 2. Oftewel, $\mathbf{Y} = \tilde{\mathbf{Y}}_{\mathbf{proj}}$ geeft $\tilde{\mathbf{Y}} = \mathbf{Y} = \tilde{\mathbf{Y}}_{\mathbf{proj}}$. Stel er bestaat een zadelpunt, dan verwachten we dus dat het $(\mathbf{Y}, \tilde{\mathbf{Y}}) = (\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$ zal zijn.

Nu willen we controleren of $(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$ inderdaad een zadelpunt is. Als $(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$ namelijk een zadelpunt is, geldt lemma 2 en geldt dus stelling 1. We beschouwen de functie uit (7).

$$\begin{aligned} f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}}) &= - \| \hat{\mathbf{Y}} - \tilde{\mathbf{Y}}_{\mathbf{proj}} \|^2 \\ f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}) &= \| \tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}_{\mathbf{proj}} \|^2 - \| \hat{\mathbf{Y}} - \tilde{\mathbf{Y}}_{\mathbf{proj}} \|^2 \end{aligned}$$

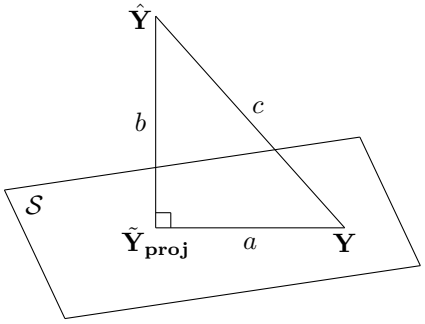
Omdat $\| \tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}_{\mathbf{proj}} \|^2 \geq 0$ geeft dit ons voorwaarde (4) van de definitie van een zadelpunt:

$$\forall \tilde{\mathbf{Y}} \in S : f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}) \geq f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}}).$$

Voor voorwaarde (3) van de definitie van een zadelpunt willen we laten zien dat geldt: $\forall \mathbf{Y} \in \mathbf{S} : \mathbf{f}(\mathbf{Y}, \tilde{\mathbf{Y}}_{\mathbf{proj}}) \leq \mathbf{f}(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$, oftewel

$$\| \tilde{\mathbf{Y}}_{\mathbf{proj}} - \mathbf{Y} \|^2 - \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2 \leq - \| \hat{\mathbf{Y}} - \tilde{\mathbf{Y}}_{\mathbf{proj}} \|^2 . \tag{8}$$

We bekijken hiervoor onze gegevens in figuur 1. $\mathbf{Y}$ en $\tilde{\mathbf{Y}}_{\mathbf{proj}}$ liggen in het vlak van consistente vectoren, S. $\tilde{\mathbf{Y}}_{\mathbf{proj}}$ hebben we gedefinieerd als $\tilde{\mathbf{Y}}_{\mathbf{proj}} = \arg\min_{\tilde{\mathbf{Y}} \in S} \| \tilde{\mathbf{Y}} - \hat{\mathbf{Y}} \|^2$; dit is dus de loodrechte projectie van $\hat{\mathbf{Y}}$ op S. We



Figuur 1: Illustratie van de stelling van Pythagoras.

hebben dus eigenlijk een driehoek met een rechte hoek. We noemen de afstand tussen $\mathbf{Y}$ en $\tilde{\mathbf{Y}}_{\mathbf{proj}}$, de afstand tussen $\tilde{\mathbf{Y}}_{\mathbf{proj}}$ en $\hat{\mathbf{Y}}$ en de afstand tussen $\hat{\mathbf{Y}}$ en $\mathbf{Y}$ respectievelijk a, b en c. (8) kunnen we dan schrijven als:

$$a^2 - c^2 \leq -b^2.$$

Deze ongelijkheid volgt uit de stelling van Pythagoras, die stelt dat in een rechthoekige driehoek met rechthoekszijden a en b en schuine zijde c: $a^2 + b^2 = c^2$.

Aan beide voorwaarden voor een zadelpunt wordt voldaan, dus $(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$ is een zadelpunt. Volgens lemma 2 geldt nu dat $\min_{\tilde{\mathbf{Y}} \in S} \max_{\mathbf{Y} \in S} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})) = \max_{\mathbf{Y} \in S} \min_{\tilde{\mathbf{Y}} \in S} ((\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})) = - \parallel \hat{\mathbf{Y}} - \tilde{\mathbf{Y}}_{\mathbf{proj}} \parallel^2 \leq 0.$ □

3 Generalisatie naar Bregman divergenties

In paragraaf 2.4 wordt bij het meten van de kwaliteit van een voorspelling gebruikgemaakt van de kwadratische Euclidische afstand. We zouden de kwaliteit ook kunnen meten met behulp van een andere afstandsmaat. De kwadratische Euclidische afstand is de bekendste maat van de groep afstandsmaten genaamd Bregman divergenties.

3.1 Definitie

Voordat Bregman divergenties gedefinieerd kunnen worden moeten we de *Legendre functies* beschouwen. We noemen een functie $f : \mathcal{A} \to \mathbb{R}$ Legendre als deze voldoet aan de volgende voorwaarden [1, section 11.2]:

1. $\mathcal{A} \subseteq \mathbb{R}^d$ is niet leeg en het inwendige van $\mathcal{A}$, $\text{int}(\mathcal{A})$, is convex;
2. f is strikt convex met continue eerste partiële afgeleiden in heel $\text{int}(\mathcal{A})$;
3. Als $x_1, x_2, \ldots \in \mathcal{A}$ een rij is die naar een randpunt van $\mathcal{A}$ convergeert, dan $\parallel \nabla f(x_n) \parallel \to \infty$ als $n \to \infty$.

De Bregman divergentie is de functie $D_f : \mathcal{A} \times \text{int}(\mathcal{A}) \to \mathbb{R}$, voortgebracht door een Legendre functie $f : \mathcal{A} \to \mathbb{R}$, gedefinieerd door:

$$D_f(x, y) = f(x) - f(y) - (x - y)\nabla f(y). \tag{9}$$

De eerder behandelde Bregman divergentie, de kwadratische Euclidische afstand, wordt bijvoorbeeld voortgebracht door de Legendre functie, gedefinieerd door $f(x) = \| x \|^2 = \langle x\,, x\rangle$. Als we deze invullen in (9) krijgen we:

$$\begin{aligned} D_f(x,y) &= \| x \|^2 - \| y \|^2 - (x-y)\nabla \| y \|^2 = \langle x\,, x\rangle - \langle y\,, y\rangle - \langle x-y\,, \nabla f(y)\rangle \\ &= \langle x\,, x\rangle - \langle y\,, y\rangle - \langle x-y\,, 2y\rangle = \langle x-y\,, x-y\rangle = \| x-y \|^2 . \end{aligned}$$

Dit geeft ons dus als functie voor de kwadratische afstand: $D_f(x,y) = \| x-y \|^2$.

Informeel meet (9) de afstand tussen een Legendre functie f en z'n raaklijn. Aangezien f strikt convex is, is deze dus nooit negatief.

3.2 Generalisatie

We willen stelling 1 generaliseren naar een stelling die geldt voor alle Bregman divergenties. ℓ beschouwen we nu niet meer als de kwadratische afstand, maar als een Bregman divergentie in het algemeen. De uitdrukking voor V, (2), beschouwen nu ook voor ℓ een Bregman divergentie. Hierbij generaliseren we ook S naar een meer algemene verzameling T.

Stelling 2. *Laat ℓ een Bregman divergentie zijn, $T \subseteq \mathbb{R}^d$ een gesloten en convexe verzameling waarvoor $\mathcal{A} \cap T = \emptyset$. Voor elke voorspelling $\hat{\mathbf{Y}} \in \text{int}(\mathcal{A})$ bestaat de unieke projectie $\tilde{\mathbf{Y}}_{\mathbf{proj}} = \arg\min_{\tilde{\mathbf{Y}} \in T} \ell(\tilde{\mathbf{Y}}, \hat{\mathbf{Y}})$ en geldt voor de waarde van V:*

$$V = -\ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}) \leq 0,$$

waarbij voor de speltheoretisch optimale voorspellingen geldt dat $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_{\mathbf{proj}}$.

Voor het bewijs van deze stelling kunnen we gebruikmaken van de onderstaande twee lemma's.

We citeren bij deze lemma 11.2 van [1]:

Lemma 3. *Voor alle Legendre functies $f : \mathcal{A} \to \mathbb{R}$, voor gesloten, convexe verzamelingen $T \subseteq \mathbb{R}^d$ zodanig dat $T \cap \mathcal{A} \neq \emptyset$ en voor alle $\mathbf{Y} \in \text{int}(\mathcal{A})$, bestaat de Bregman projectie van $\mathbf{Y}$ op T, $\tilde{\mathbf{Y}}_{\mathbf{proj}} = \arg\min_{\tilde{\mathbf{Y}} \in T} \ell(\tilde{\mathbf{Y}}, \mathbf{Y})$, en is deze uniek.*

Lemma 4 (Gegeneraliseerde ongelijkheid van Pythagoras)**.** *Laat f een Legendre functie zijn. Voor alle $\hat{\mathbf{Y}} \in \text{int}(\mathcal{A})$ en voor alle convexe, gesloten verzamelingen $T \subseteq \mathbb{R}^d$, waarbij $T \cap \mathcal{A} \neq \emptyset$, als $\tilde{\mathbf{Y}}_{\mathbf{proj}} = \arg\min_{\mathbf{X} \in T \cap \mathcal{A}} D_f(\mathbf{X}, \hat{\mathbf{Y}})$, geldt:*

$$\forall \mathbf{Y} \in T : D_f(\mathbf{Y}, \hat{\mathbf{Y}}) \geq D_f(\mathbf{Y}, \tilde{\mathbf{Y}}_{\mathbf{proj}}) + D_f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}).$$

Voor dit lemma en het bewijs hiervan refereren we naar lemma 11.3 van [1].

Bewijs stelling 2. T is per definitie een gesloten en convexe verzameling en er geldt $T \subseteq \mathbb{R}^d$ zodat $\mathcal{A} \cap T \neq \emptyset$. Wegens lemma 3 bestaat nu de projectie $\tilde{\mathbf{Y}}_{\mathbf{proj}} = \arg\min_{\tilde{\mathbf{Y}} \in T} \ell(\tilde{\mathbf{Y}}, \hat{\mathbf{Y}})$ en is deze uniek.
Stel nu dat (5) geldt voor de functie

$$f(\mathbf{Y}, \tilde{\mathbf{Y}}) = \ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}}), \tag{10}$$

waarbij $\hat{\mathbf{Y}}$ vast, aangezien dit de oorspronkelijke voorspellingen zijn:

$$\min_{\tilde{\mathbf{Y}} \in T} \max_{\mathbf{Y} \in T} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})) = \max_{\mathbf{Y} \in T} \min_{\tilde{\mathbf{Y}} \in T} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})).$$

Deze laatste uitdrukking kunnen we herschrijven:

$$\begin{aligned} \max_{\mathbf{Y} \in T} \min_{\tilde{\mathbf{Y}} \in T} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})) &\overset{(\star)}{=} \max_{\mathbf{Y} \in T} (-\ell(\mathbf{Y}, \hat{\mathbf{Y}})) \\ = -\min_{\mathbf{Y} \in T} \ell(\mathbf{Y}, \hat{\mathbf{Y}}) &= -\ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}) \overset{(\diamond)}{\leq} 0, \end{aligned}$$

waarbij $\tilde{\mathbf{Y}}_{\mathbf{proj}}$ zoals in stelling 2.
$(\star)$ geldt aangezien $\ell(\mathbf{Y}, \tilde{\mathbf{Y}})$ van de vorm $f(\mathbf{Y}) - f(\tilde{\mathbf{Y}}) - (\mathbf{Y} - \tilde{\mathbf{Y}}) \nabla f(\tilde{\mathbf{Y}})$ is, waarvoor vanwege convexiteit gold dat dit nooit negatief kan worden, maar wel 0 kan worden. Voor minimalisatie kiezen we $\tilde{\mathbf{Y}}$ zó dat $\ell(\mathbf{Y}, \tilde{\mathbf{Y}})$ gelijk aan 0 wordt, namelijk $\tilde{\mathbf{Y}} = \mathbf{Y}$. Ook $(\diamond)$ geldt wegens ditzelfde argument: $\ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}) \geq 0$.

We gaan weer de stappen van bovenstaande omschrijving vertalen naar zetten van de spelers. Speler 1 kiest om (10) te maximaliseren $\mathbf{Y} = \tilde{\mathbf{Y}}_{\mathbf{proj}}$. Speler 2 kiest voor minimalisatie $\tilde{\mathbf{Y}} = \mathbf{Y}$. Uit de zet van speler 1, volgt de keuze voor de zet van speler 2. Oftewel, uit $\mathbf{Y} = \tilde{\mathbf{Y}}_{\mathbf{proj}}$ volgt $\tilde{\mathbf{Y}} = \mathbf{Y} = \tilde{\mathbf{Y}}_{\mathbf{proj}}$. Stel er bestaat een zadelpunt, dan verwachten we dus dat het $(\mathbf{Y}, \tilde{\mathbf{Y}}) = (\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$ zal zijn.

We gaan controleren of dit een zadelpunt is. Voldoet $(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$ aan voorwaarden (3) en (4)? Als $(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$ namelijk een zadelpunt is, geldt lemma 2 en geldt dus stelling 2.
We beschouwen de functie uit (10).

$$\begin{aligned} f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}}) &= -\ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}) \\ f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}) &= \ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}) - \ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}) \end{aligned}$$

Aangezien $\ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}})$ van de vorm (9) is, waarvoor gold dat het niet negatief kan zijn vanwege convexiteit, geldt:

$$\forall \tilde{\mathbf{Y}} \in T : f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}) \geq f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}}).$$

Aan voorwaarde (4) wordt dus voldaan. Nu vragen we ons nog af of ook aan voorwaarde (3) wordt voldaan. Hiervoor zou moeten gelden:

$$\forall \mathbf{Y} \in T : f(\mathbf{Y}, \tilde{\mathbf{Y}}_{\mathbf{proj}}) \leq f(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}}),$$

oftewel:

$$\ell(\mathbf{Y}, \tilde{\mathbf{Y}}_{\mathbf{proj}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}}) \leq -\ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}). \tag{11}$$

Voor deze voorwaarde kunnen we gebruikmaken van lemma 4. Voor T gold dat deze gesloten en convex is en deze hadden we zo gekozen dat $T \cap A \neq \emptyset$. Volgens lemma 4 geldt nu:

$$\forall \mathbf{Y} \in T : \ell(\mathbf{Y}, \hat{\mathbf{Y}}) \geq \ell(\mathbf{Y}, \tilde{\mathbf{Y}}_{\mathbf{proj}}) + \ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}).$$

Hieruit volgt de benodigde voorwaarde (11).

Aan beide voorwaarden voor het zadelpunt wordt voldaan, dus $(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \tilde{\mathbf{Y}}_{\mathbf{proj}})$ is een zadelpunt. Volgens lemma 2 geldt nu dat

$$\begin{aligned}\min_{\tilde{\mathbf{Y}} \in T} \max_{\mathbf{Y} \in T} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})) &= \max_{\mathbf{Y} \in T} \min_{\tilde{\mathbf{Y}} \in T} (\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})) \\ &= -\ell(\tilde{\mathbf{Y}}_{\mathbf{proj}}, \hat{\mathbf{Y}}) \leq 0.\end{aligned}$$

□

Voor alle Bregman divergenties geldt dus dat kwaliteit van de nieuwe, consistente voorspellingen altijd minstens even hoog is als de kwaliteit van de oorspronkelijke, inconsistente voorspellingen.

3.3 Voorbeeld kwadratische Euclidische afstand

We zullen laten zien dat uit stelling 2 inderdaad stelling 1 volgt.

Kies $\mathcal{A} = \mathbb{R}^d$ en neem T gelijk aan de verzameling van consistente vectoren, dus $T = S$, zodat $\mathcal{A} \cap S \neq \emptyset$. We hoeven nu alleen nog aan te tonen dat S inderdaad gesloten en convex is. Hiervoor zullen we eerst geslotenheid en convexiteit definiëren.

Definitie 2. *Een verzameling $X \subseteq \mathbb{R}^d$ heet* gesloten *als voor elke rij $(x_n)_{n=0}^{\infty} \subseteq X$ die convergeert naar $x \in \mathbb{R}^d$ geldt dat $x \in X$.*

Definitie 3. *Een verzameling $X \subseteq \mathbb{R}^d$ heet* convex *als voor alle $x, y \in X$ en voor alle $t \in [0, 1]$ geldt dat $tx + (1-t)y \in X$.*

We zullen allereerst laten zien dat S gesloten is.
Laat

$$(x^{(k)})_{k=0}^{\infty} = (x_1^{(k)}, x_2^{(k)}, \ldots, k_d^{(k)})_{k=0}^{\infty} \subseteq S$$

convergeren in $\mathbb{R}^d$ naar $x = (x_1, \ldots, x_d)$. Voor elke k geldt:

$$\sum_{i=1}^{d-1} x_i^{(k)} = x_d^{(k)}.$$

Dan geldt dat $x_d = \lim_{k\to\infty} x_d^{(k)} = \lim_{k\to\infty} \sum_{i=1}^{d-1} x_i^{(k)} = \sum_{i=1}^{d-1} \lim_{k\to\infty} x_i^{(k)} = \sum_{i=1}^{d-1} x_i$, dus $x \in S$. S is een gesloten verzameling.

Dan zullen we nog laten zien dat S ook convex is.
We nemen twee elementen uit S: $(x_1, \ldots, x_d), (y_1, \ldots, y_d) \in S$. Dan geldt voor alle $t \in [0, 1]$:

$$\begin{aligned}t \cdot (x_1, \ldots, x_d) + (1-t) \cdot (y_1, \ldots, y_d) &= (tx_1 + (1-t)y_1, \ldots, tx_d + (1-t)y_d) \\ &=: (z_1, \ldots, z_d),\end{aligned}$$

en

$$\sum_{i=1}^{d-1} z_i = \sum_{i=1}^{d-1} (tx_i + (1-t)y_i) = t \sum_{i=1}^{d-1} x_i + (1-t) \sum_{i=1}^{d-1} y_i = tx_d + (1-t)y_d = z_d,$$

omdat $x, y \in S$, dus z $\in S$. S is dus ook een convexe verzameling.

Dit is een voorbeeld van hoe het bewijs van stelling 2 inderdaad toegespitst kan worden voor het bewijs van de stelling voor een specifieke Bregman divergentie, in dit geval de kwadratische Euclidische afstand.

4 Toepassing op Kullback-Leibler divergentie

Nu we weten dat de methode voor alle Bregman divergenties werkt, kunnen we naar de toepassing van een andere Bregman divergentie in het bijzonder kijken. We richten ons in dit hoofdstuk op de Kullback-Leibler divergentie.

4.1 Definitie

De Kullback-Leibler divergentie is interessant om ons op te focussen, aangezien deze het verschil meet tussen twee kansverdelingen. De Kullback-Leibler divergentie tussen kansverdelingen $\mathbf{p}$ en $\mathbf{q}$ wordt gedefinieerd door:

$$D_f(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{d} p_i \log_2 \frac{p_i}{q_i}. \tag{12}$$

Als $\mathbf{p}$ een discrete kansverdeling is, met $\sum_{i=1}^{d} p_i = 1$, $p_i \geq 0$, is de negatieve entropie $f(\mathbf{p}) = \sum_{i=1}^{d} p_i \log_2 p_i$ de voortbrengende, convexe functie. Voor de afleiding van (12) bekijken we eerst de gradiënt van f op $\mathbf{q}$:

$$\begin{aligned} f(\mathbf{q}) &= \sum_{i=1}^{d} q_i \log_2 q_i = \sum_{i=1}^{d} q_i \cdot \frac{\ln q_i}{\ln 2} \\ \nabla f(\mathbf{q}) &= \sum_{i=1}^{d} \left(\frac{\ln q_i}{\ln 2} + q_i \cdot \frac{\frac{1}{q_i}}{\ln 2} \right) = \sum_{i=1}^{d} \left(\log_2 q_i + \frac{1}{\ln 2} \right) \\ &= \sum_{i=1}^{d} (\log_2 q_i + \log_2 e). \end{aligned}$$

Dus voor de bijbehorende Bregman divergentie geldt dan, door (9) in te vullen:

$$\begin{aligned} D_f(\mathbf{p}, \mathbf{q}) &= \sum_{i=1}^{d} p_i \log_2 p_i - \sum_{i=1}^{d} q_i \log_2 q_i - \sum_{i=1}^{d} (p_i - q_i) \nabla f(\mathbf{q}) \\ &= \sum_{i=1}^{d} p_i \log_2 p_i - \sum_{i=1}^{d} q_i \log_2 q_i - \sum_{i=1}^{d} (p_i - q_i)(\log_2 q_i + \log_2 e) \\ &= \sum_{i=1}^{d} p_i \log_2 \frac{p_i}{q_i} - \log_2 e \sum_{i=1}^{d} (p_i - q_i) \\ &= \sum_{i=1}^{d} p_i \log_2 \frac{p_i}{q_i}, \end{aligned}$$

waarbij de laatste term wegvalt aangezien $\sum_{i=1}^{d} (p_i - q_i) = 1 - 1 = 0$.

4.2 Toepassing

Nu we de Kullback-Leibler divergentie willen gebruiken, zullen we als input kansverdelingen nodig hebben. Zowel de originele, inconsistente voorspellingen $\hat{\mathbf{Y}}$ als de nieuwe, consistente voorspellingen $\tilde{\mathbf{Y}}$ én het daadwerkelijke gebruik $\mathbf{Y}$ zullen we in termen van kansverdelingen moeten beschrijven. De kansverdelingen van de originele en nieuwe voorspellingen noemen we respectievelijk $\hat{\mathbf{p}}$ en $\tilde{\mathbf{p}}$. $\mathbf{Y}$ is echter een vector van data, geen kansverdeling. Willen we gebruik kunnen maken van de Kullback-Leibler divergentie, zullen we $\mathbf{Y}$ dus moeten transformeren naar een kansverdeling, zeg $\mathbf{p}^{\mathbf{Y}}$. Dit doen we door de kansverdeling te nemen die kans 1 heeft op de uitkomst $\mathbf{Y}$ en kans 0 heeft op iedere andere uitkomst.

Aangezien ℓ alleen gedefinieerd is voor eindige dimensies zijn we voor de Kullback-Leibler divergentie dus beperkt tot eindig dimensionale vectoren en daarmee tot eindig dimensionale kansverdelingen. Daarom definiëren we allereerst de onderliggende verzameling $\mathcal{Y}$ als een eindige verzameling van vectoren. Het is bijvoorbeeld voldoende als alle $\mathbf{y} \in \mathcal{Y}$ in hetzelfde aantal decimalen geschreven kunnen worden; dit is een realistische eis als we bedenken dat bijvoorbeeld op computers ook alles in hetzelfde aantal bits beschreven moet kunnen worden. Vervolgens definiëren we $\mathcal{A}$ als de verzameling van alle kansverdelingen $\mathbf{p}$ op $\mathcal{Y}$, oftewel $\mathcal{A} = \{p | p_i \geq 0, \sum_{i=1}^{d} p_i = 1\}$. Omdat $\{\mathbf{p}^{\mathbf{Y}}\}$ niet convex is kiezen we de kleinste convexe set die $\{\mathbf{p}^{\mathbf{Y}}\}$ bevat. We nemen $T = \{\mathbf{p} | \mathbf{p}(S) = 1\}$, oftewel T gelijk aan alle kansverdelingen waarvoor geldt dat de kans op een consistente vector gelijk aan 1 is.

Ons oorspronkelijke probleem gaat over kansverdelingen en data. Hier zouden we graag iets over willen zeggen door middel van een stelling. Omdat we hier stelling 2 niet direct voor kunnen gebruiken, gaan we dit probleem van boven begrenzen door een probleem waar we wel stelling 2 op kunnen toepassen. Als we namelijk voor een bovengrens hebben aangetoond dat hij nooit een positieve waarde kan aannemen, geldt dat ook voor het daadwerkelijke probleem.

Uit de toepassing van stelling 2 op de gebruikte bovengrens volgt een mogelijke goede projectie, $\tilde{\mathbf{p}}_{\mathbf{proj}} = \arg\min_{\tilde{\mathbf{p}} \in T} D_f(\tilde{\mathbf{p}}, \hat{\mathbf{p}})$, waar we dan ook gebruik van zullen maken in de volgende stelling.

Stelling 3. *Laat D_f de Kullback-Leibler divergentie zijn, laat $\mathcal{A} = \{p | p_i \geq 0, \sum_{i=1}^{d} p_i = 1\}$ en $T = \{\mathbf{p} | \mathbf{p}(S) = 1\}$. Voor elke voorspelling $\hat{\mathbf{p}} \in \text{int}(\mathcal{A})$ bestaat de unieke projectie $\tilde{\mathbf{p}}_{\mathbf{proj}} = \arg\min_{\tilde{\mathbf{p}} \in T} D_f(\tilde{\mathbf{p}}, \hat{\mathbf{p}})$ en voor de waarde van*

$$U = \max_{\mathbf{Y} \in S} (D_f(\mathbf{p}^{\mathbf{Y}}, \tilde{\mathbf{p}}_{\mathbf{proj}}) - D_f(\mathbf{p}^{\mathbf{Y}}, \hat{\mathbf{p}})) = \max_{\mathbf{Y} \in S} \left(\log_2 \frac{1}{\tilde{\mathbf{p}}_{\mathbf{proj}}(\mathbf{Y})} - \log_2 \frac{1}{\hat{\mathbf{p}}(\mathbf{Y})} \right)$$

geldt:

$$U \leq -D_f(\tilde{\mathbf{p}}_{\mathbf{proj}}, \hat{\mathbf{p}}) \leq 0.$$

Voorbeeld 3 uit [2] geeft ons dat $\tilde{\mathbf{p}}_{\mathbf{proj}}(\mathbf{Y}) = \hat{\mathbf{p}}(\mathbf{Y} | \mathcal{S})$.

Voor het bewijs van stelling 3 maken we gebruik van het feit dat $T = \{\mathbf{p} | \mathbf{p}(S) = 1\}$ gesloten en convex is. Dit zal allereerst worden aangetoond.

Lemma 5. *T is een gesloten verzameling.*

Bewijs. Laat $(p_k)_{k=0}^{\infty} \subseteq T$ convergeren naar $\mathbf{p}$. Dan geldt:

$$\mathbf{p}(S) = \lim_{k\to\infty} p_k(S) = \lim_{k\to\infty} 1 = 1,$$

dus $\mathbf{p}(S) \in T$. □

Lemma 6. *T is een convexe verzameling.*

Bewijs. Neem $\mathbf{p}, \mathbf{q} \in T$ en $t \in [0, 1]$. Dan geldt:

$$(t\mathbf{p} + (1-t)\mathbf{q})(S) = t\mathbf{p}(S) + (1-t)\mathbf{q}(S) = t + 1 - t = 1,$$

dus $t\mathbf{p} + (1-t)\mathbf{q} \in T$. □

Bewijs stelling 3. We passen stelling 2 toe op de Kullback-Leibler divergentie op kansverdelingen $\tilde{\mathbf{p}}, \mathbf{p} \in T$, waarbij $T = \{\mathbf{p} | \mathbf{p}(S) = 1\}$. In dit geval geldt voor (2):

$$U' = \min_{\tilde{\mathbf{p}}\in T} \max_{\mathbf{p}\in T} D_f(\mathbf{p}, \tilde{\mathbf{p}}) - D_f(\mathbf{p}, \hat{\mathbf{p}}),$$

en uit stelling 2 volgt nu voor de waarde van U':

$$U' = -D_f(\tilde{\mathbf{p}}_{\mathbf{proj}}, \hat{\mathbf{p}}) \leq 0.$$

Aangezien T alle $\mathbf{p}^{\mathbf{Y}}, \mathbf{Y} \in S$ bevat, plus de convexe combinaties hiervan, geldt $\{\mathbf{p}^{\mathbf{Y}} | \mathbf{Y} \in S\} \subseteq T$. Hieruit volgt dat $U \leq U'$. Conclusie: $U \leq U' \leq 0$. □

In het geval van de Kullback-Leibler blijkt dus inderdaad ook dat de kwaliteit van de nieuwe, consistente voorspellingen altijd minstens even hoog is als bij de oorspronkelijke voorspellingen.

5 Samenvatting en toekomstig onderzoek

5.1 Samenvatting

In deze scriptie hebben we ons beziggehouden met het maken van bruikbare, nauwkeurige voorspellingen voor elektriciteitsverbruik. Allereerst is verklaard wanneer een voorspelling in de praktijk bruikbaar is, namelijk wanneer deze consistent is. In eerste instantie hebben we de kwadratische afstand als afstandsmaat voor de kwaliteit van de voorspellingen gebruikt. Bij het omzetten van inconsistente naar consistente voorspellingen mocht de kwaliteit nooit verslechteren. Daarom moest het verlies tussen het daadwerkelijke gebruik en de nieuwe, consistente voorspelling altijd kleiner zijn dan het verlies tussen het daadwerkelijke gebruik en de originele voorspelling. Hiervoor hebben we een speltheoretische optimale methode uiteengezet.

In stelling 1 hebben we, gebaseerd op theorie 1 van [5], laten zien dat voor de specifieke Bregman divergentie, de kwadratische Euclidische afstand, een unieke projectie voor de voorspelling bestaat, die ons verzekert dat de kwaliteit van deze nieuwe, consistente voorspelling nooit verslechtert. Vervolgens hebben we deze stelling gegeneraliseerd naar de algemene stelling 2, waarin we

hebben aangetoond dat dit argument voor alle Bregman divergenties opgaat. Tot slot is hiervandaan gespecialiseerd naar één specifieke Bregman divergentie, de Kullback-Leibler divergentie. In stelling 3 bewijzen we dat voor deze divergentie, die het verschil tussen twee kansverdelingen meet, ook zo'n unieke projectie bestaat, waarmee de kwaliteit van de nieuwe, consistente voorspelling altijd minstens even goed is als de kwaliteit van de originele. Terwijl het bewijs van stelling 2 analoog was aan het bewijs van stelling 1, was het bewijs van stelling 3 van een andere structuur. Hierbij konden we de algemene stelling 2 niet toepassen op ons oorspronkelijke probleem en hebben daarom een bovengrens gebruikt waarop dit wel mogelijk was.

5.2 Toekomstig onderzoek

Ondanks stelling 2 en stelling 3 als nieuwe resultaten, valt er nog veel te onderzoeken. In paragraaf 4.2 beperken we ons bijvoorbeeld tot $\mathcal{Y}$ als eindige verzameling van vectoren, wat uitgebreid kan worden naar een onderzoek betreffende $\mathcal{Y}$ als continue verzameling van vectoren. Ook hiervoor, wanneer kansen vervangen worden door kansdichtheden, is de verwachting dat er een unieke projectie zal bestaan, die het gewenste resultaat zal leveren.

In dezelfde paragraaf stellen we, op basis van een bovengrens van U, iets over de waarde van U. Een aanvulling hierop zou een daadwerkelijke uitschrijving, zoals bijvoorbeeld in paragraaf 2.5 wordt gedaan, van U zijn, om hier direct een exacte uitdrukking voor te vinden.

Naast de Kullback-Leibler divergentie, is er nog een aantal Bregman divergenties dat interessant is om op in te zoomen. Als uitbreiding van deze scriptie zou verder bijvoorbeeld nog naar de veelgebruikte Itakura-Saito divergentie, de Bregman divergentie die het verschil tussen twee spectra meet, gekeken kunnen worden.

Referenties

[1] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[2] P. Harremoës. Information topologies with applications. In *Entropy, Search, Complexity*. Springer Berlin Heidelberg, 2007.

[3] R. J. Hyndman, R. Ahmed, G. Athanasopoulos, and H. Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis*, 55:2579–2589, 2011.

[4] R. Stone, D. Champernowne, and J. Meade. The precision of national income estimates. *The Review of Economic Studies*, 9(2):111–125, 1942.

[5] T. van Erven and J. Cugliari. Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. *Modeling and Stochastic Learning for Forecasting in High Dimensions*, pages 297–317, 2015.