BSC PROJECT: SCHATTEN VAN MONTE CARLO CONVERGENTIE

JORIS BIERKENS

In wiskundige en andere exacte disciplines komen vaak ingewikkelde kansverdelingen om de hoek kijken. Dit is niet alleen zo bij kansrekening en statistiek, maar ook in de natuurkunde (*statistische fysica*), optimalisatie (*simulated annealing*) en machine learning (*deep learning*). In de praktijk zijn kansverdelingen vaak lastige structuren om mee te werken. Een methode om praktische berekeningen uit te voeren is de *Monte Carlo* methode. Hierbij wordt een kansverdeling π op $\mathbb{R}^d$ benaderd door een groot aantal waarden $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ in $\mathbb{R}^d$. Deze waarden representeren trekkingen van de kansverdeling π. Met behulp van de trekkingen $(\boldsymbol{x}^{(i)})$ kun je nu praten over een benaderende (empirische) kansverdeling $\hat{\pi}_n := \frac{1}{n} \sum_{i=1}^n \delta_{\boldsymbol{x}^{(i)}}$ waar $\delta_{\boldsymbol{z}}$ de Dirac delta-'functie' in $\boldsymbol{z}$ voorstelt. De wet van de grote aantallen zegt dat onder bepaalde voorwaarden (welke?) de empirische verwachting convergeert:

$$\mathbb{E}_{\hat{\pi}_n} f(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^n f(\boldsymbol{x}^{(i)}) \to \mathbb{E}_\pi f(\boldsymbol{X}) \quad \text{met kans 1 als } n \to \infty$$

voor een ruime klasse van functies $f : \mathbb{R}^d \to \mathbb{R}$, en deze convergentie kun je opvatten als het doel van een Monte Carlo algoritme.

Om te kunnen bepalen of een Monte Carlo algoritme goed werkt, moeten we kunnen vaststellen wat de kwaliteit van de gevonden benadering is op basis van de trekkingen $(\boldsymbol{x}^{(i)})$. Een complicatie hierbij is dat we er niet vanuit kunnen gaan dat we $\mathbb{E}_\pi f(\boldsymbol{X})$ kennen. De onderzoeksvraag van dit project is daarom:

> *Hoe kunnen we voor een functie $f : \mathbb{R}^d \to \mathbb{R}$ bepalen wat de benaderingsfout $e(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}) := |\mathbb{E}_{\hat{\pi}_n} f(\boldsymbol{X}) - \mathbb{E}_\pi f(\boldsymbol{X})|$ is, zonder $\mathbb{E}_\pi f(\boldsymbol{X})$ te kennen?*

Dit statistische vraagstuk heeft een zeer groot praktisch nut voor alle bovengenoemde wetenschapsgebieden. De schattingsmethoden worden namelijk zowel gebruikt bij de ontwikkeling van nieuwe algoritmes, als om te besluiten wanneer een lopend algoritme stopgezet kan worden. Om te beginnen zullen we kijken naar enkele bekende methodes, zoals de *batch means* methode om de *asymptotische variantie* te schatten: dit is een getal σ_f^2 zodat

$$\sqrt{n}(\mathbb{E}_{\hat{\pi}^n} f(\boldsymbol{X}) - \mathbb{E}_\pi f(\boldsymbol{X})) \sim \mathcal{N}(0, \sigma_f^2), \text{bij benadering voor grote } n,$$

ofwel met andere woorden, $e(\boldsymbol{x}^(1), \ldots, \boldsymbol{x}^{(n)}) \approx \sigma_f / \sqrt{n}$. Na een eerste verkenning van deze relatief bekende methode zul je in de literatuur op zoek gaan naar alternatieven. Zo zul je in aanraking komen met de nieuwste wetenschappelijke inzichten in dit veld en wellicht lukt het je ook om zelf een nieuw idee bij te dragen. Afhankelijk van je interesse zul je dit onderwerp theoretisch benaderen, met de nadruk op het wiskundig precies maken van de gebruikte schattingen en benaderingen, of je concentreren op het implementeren en testen van de schattingsmethoden. Een combinatie van beide is ook zeker mogelijk. Om alvast iets meer te lezen verwijs ik je naar [RC04, Hoofdstuk 12] (toegepast) of [vdV10, §5.1] (meer theoretisch).

Referenties

[RC04] Christian P Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

[vdV10] A W van der Vaart. Time Series. *Leiden University Lecture Notes*, pages 1–232, 2010.