PLEASE NOTE: **My interests are not limited only to this project. This proposal shows a subset of my research interests and skills.**

Conduct a study on gradient estimators aiming at increasing their generality and performance. Use case: Annotation/Labeling/Rating Systems

January 4, 2023

Abstract

This paper presents an attempt on developing a framework of gradient estimators (GE) that can be exploited to efficiently learn models for an as-wider-as-possibly range of problems.

In the Introduction section, this proposal presents a literature review of the work done by the most eminent researchers with regard to GE; their pros and cons; research related to plausibility, i.e. the difference between explanations generated from models and those generated by humans when using different GE. It also highlights their disadvantages and limits in addressing certain challenges.

Later in this section, it is briefly presented an examination of a novel family of GE called (A)IMLE which can be applied to hybrid models, i.e. models that mix neural networks with a discrete component. This gradient estimator family has demonstrated competitive performance with respect to other estimators and in particular supremacy with regard to precision. Just being competitive with other estimators is another advantage of this class of GE thanks to its generality and the limitations that other estimators present. This is a source of inspiration urging us to further investigate and analyze this novelty and provides room for new inventions. This is the first direction.

The second direction of this study is to analyze the reasons behind plausibility issues related to Rating systems. The two directions are brought together in one project because rating systems constitute one of the realms where gradient estimators are widely used and can be compared with each other with respect to their performance.

Objectives and goals are formulated in the section after the Introduction. The future work will cover an examination of the AIMLE in other tasks.

Keywords: Probabilistic Machine Learning, CNN, NLP, Explainable AI, Interpretability, Annotations, Explanations, Highlights, Model Plausibility, Model Fidelity, Gradient Estimators (GE), Discrete Latent Variable, Combinatorial Optimization (CO), Discrete Exponential Family of Distributions, Implicit Finite Differentiation via Perturbation, Neural Networks(NN), Implicit Maximum Likelihood Estimation (IMLE), Adaptive IMLE (AIMLE), Variational Inference, Mean Squared Error(MSE), Rating System (RS), (Visual-Textual) Entailment, Natural Language Explanations/Rationals, Multi-Aspect Review, Image Features extract

1 Introduction

1.1 Importance of gradient estimation and explainable AI

The magic power of solving a large aptitude of problems emerging from a high variety of fields through ML systems it is not only intriguing but also of vital interest due to unprecedentedly fast solutions they yield for these problems, which sometimes, due to their complexity, without automation, could have been impossible to be solved in time or even to be tackled at all like medicine, robotics, chatbots, finance, space exploration, defense sector, etc..

Although the achievements of AI are impressive, there is, of course, a lot to be done not only in the direction of attacking unsolved problems but as well in the direction of improving the performance of AI systems, especially with respect to accuracy and time efficiency. Along with the increasing demand for efficient AI solutions, it has increased the exigency for a higher-level explanation of the AI system outcomes. The insights are not only used to add more supervision to model training, but also to give an account for the system's decisions (Wiegreffe and Marasović, 2021).

An approach tightly related to learning models consists of gradient-based methods. For well-known reasons, like the lack of differentiability or the vanishing gradient problem, it is impossible to differentiate during back-propagation. The issue becomes more interesting and intricate in the case of hybrid models, i.e when a discrete component is mixed with NN.

1.2 Alternative Solutions for Gradient Estimation

A remedy to handle these issues for discrete exponential family distributions becomes the Adaptive Implicit Maximum Likelihood Estimation (AIMLE) invented by (Minervini et al., 2022).

In fact, this is not the only solution for the gradient estimation of discrete random variables. There are known solutions (continuous relaxation) that rely on producing approximated differentiation of sampling from variables that model categorical distributions like Gumbel-Softmax trick (Maddison et al., 2016; Paulus et al., 2020; Jang et al., 2016). While for more complex models there are needed tailored solutions that work well only for specific given models (Kim et al., 2016; Grover et al., 2019; Tucker et al., 2017; Minervini et al., 2022).

"AIMLE and IMLE provide a general-purpose framework that does not require access to the linear constraints and the corresponding integer polytope C." cites Minervini et al. (2022).

1.3 How AIMLE works

How the IMLE works. Let us assume we have an end-to-end learning model where is obtained during the training of the NN component and z is the latent discrete variable distributed over exponential model $p(z;\theta)$, then the objective and the gradient to it are expressed as following:

I $L(x,y;\omega) \overset{def}{=} E_{z\sim p(z;\theta)}[\ell(f_u(z),y)]$, where x is the very first (features) input of the pipeline and y is the target output.

II $\omega=(v,u)$; $\theta=h_v(x)$; v, u are parameters learned from the first and last neural network of the pipeline respectively.

III $z\sim p(z;\theta)$ where $p(z;\theta)$ is a distribution from the discrete exponential family.

IV At a point, it is required the value of $\nabla_\theta L(x,y;\omega)=\nabla_\theta E_{z\sim p(z;\theta)}[...f()...]=\nabla_\theta E_{z\sim p(z;\theta)}[func2(z)]=\nabla_\theta E_{z\sim p(z;\theta)}[func2(\theta)]$

During the model learning while applying gradient descent to minimize (I) it amounts to taking derivatives with respect to θ of the expected loss (IV) which is the function of the function of z. Sampling latent discrete z from the distribution is one of the key challenges (intractability).

Remedy: The IMLE estimates the gradient by combining different techniques like perturb-and-MAP to sample from the $p(z;\theta)$ as described by Papandreou and Yuille (2011). As shown by Niepert et al. (2021), this sampling is equivalent to obtaining MAP states from a perturbed model under certain conditions. IMLE, in order to compute an approximation of gradient, "puts in the game" at every update step a target distribution q created as a result of perturbing the parameters of the original distribution p by a measure proportional to the gradient of the loss with respect to the discrete z computed during forward pass. An efficient surrogate loss function relies on these two distributions and its gradient is equivalent to the gradient of the KL divergence between p and q.

After applying implicit differentiation by perturbation of (Domke, 2010) and several tricks by authors, (IV) can be estimated via

$$E_{\epsilon\sim\rho(\epsilon;\theta)}[\lim_{\lambda\to 0}\frac{1}{\lambda}\{z-MAP(\ \theta+\epsilon-\lambda\nabla_z f(z))]$$

The estimator above is Adaptive IMLE (AIMLE), the extension of IMLE, which tackles the bias problem associated with IMLE.

The expectation is taken over $\epsilon\sim\rho(\epsilon)$ which is a noise distribution. Here the authors come up with a new noise distribution called Sum-of-Gamma (SOG). (A)IMLE is compared against Score Function Estimator(STE) and Straight-Through Estimator (SFE). The experiments showed that any of the three GE (AIMLE, SFE, STE) gave better performance when using SoG perturbations instead of the $Gumbel(0,\tau)$ perturbations (perturbation=perturbing every component of θ - vector of real numbers); and, that (A)IMLE-SoG is the best the combination especially when we speak about precision. AIMLE compared to IMLE requires less number range of samples for even better performance.

In case of SFE: $\nabla_\theta E_{z\sim p(z;\theta)}[f(z)]=E_{z\sim p(z;\theta)}[\nabla_\theta \log p(z;\theta)*f(z)]$ where $\nabla_\theta \log p(z;\theta)$ is called score function and $f(z)$ is the cost function.

From the other side, it emerges as a duty to further experiment and analyze a few parameters related to this distribution under different settings of the whole ML system to observe their effects on the learning of different models.

A larger means a larger bias and less sparse gradient estimator. Smaller means less bias and a more sparse gradient estimator. Too small means zero gradients; zero is back-propagated; as a result, no learning occurs. Tradeoff? In (Minervini et al., 2022)it is shown that depends on magnitude; thus normalization of the perturbation strength is imperative. The authors show how is related to other parameters and how it is learned (adapted) during the training to reach optimal tradeoff. Looking at the possibility of self-adaption of the parameters instead of human-tuning of them is **a good space worth working in.**

Niculae et al. (2018) presents another solution based on an approximation that tackles the sparsity problem for the same models class, but it takes for granted the computation of MAP. This calculation as we have mentioned above is a serious challenge.

While in the case of AIMLE, as f is a function of marginals for the complex distributions and their computation is intractable in most of the real scenarios, they are estimated as approximations of Perturb-and-MAP samples.

1.4 Advantages of AIMLE

Compared to other methods, the advantage of AIMLE is that it considers the most probable samples' and does not need optimal solutions for supervision when the discrete component is a CO algorithm because it uses the implicitly generated target distributions q. It amounts to higher precision while having comparable MSE to other competitors. Adaptive lambda which trades off between bias and gradient sparsity is learned during the training. A key advantage is its generality for discrete exponential distribution models

1.5 Challenges of AIMLE

In the discussion above, generality is described as a top advantage. But what is meant by generality? Is it worth considering the bridge between the generality in theory and its performance in practice?
In fact, the idea of producing a general-purpose family of GE is quite appealing as well the ambitious work done by authors so far. Answering the first question, we understand that we take under consideration the discrete exponential family of distributions for which fortunately does exist a closed-form expression (**?**) and where additionally in the case of AIMLE, random variable z is discrete and from an integral polytope. In a few words, we have both constraints and generality within these limits. **Although there are limits, this generality is big enough for causing problems.**

The main problem is that while saving this generality, the performance of the model to be learned must be competitive and even better than competitors. In order to check or/and achieve this, empirical and analysis support must be provided. This requires checking how well AIMLE works with different categories of tasks. Each task may bring its unique characteristics which may reflect the need for further enhancement to the whole set of rules for GE. In return, this enhancement could increase the performance on other tasks. But this needs to be checked. Side effects on other tasks are possible. Authors have picked up some tasks and provided comparative statistics on their papers which show higher precision and very similar MSE compared to other approaches which target only one specific task. This is positive, but I think that **there is room for additional realms where to check how AIMLE** performs and eventually provide analytic proof of why AIMLE is better with respect to a set of performance requirements.

1.6 Annotation/Labeling/Rating Systems

Considering the fact that the novel framework AIMLE, apart from competitive MSE, yields higher precision than other GE, it sounds reasonable to exploit its performance advantages for recommending system tasks as well.

A Rating System project presented under the section Learning To Explain in Niepert et al. (2021); Minervini et al. (2022) is among the tasks where (A)IMLE is applied. It can be noticed that models are trained separately for each aspect (appearance, smell, aroma, palate). **This approach may ignore correlations between annotations of different aspects.**

The provided experiments show for this task the same performance advantages, but there is still a lot more to be seen with regard to this task, especially accuracy. Sometimes, the model does not select some words as the human does to explain ratings(labels). The (A)IMLE attacks the generation of explanations by the model as a task of learning k-subset distributions. The performance of computing MAPs is great; it is linear in k. But **providing to system a fixed limit on the length of the explanation as a hyper-parameter k can cause problems of sufficiency.**

This project considers models that in addition to labels generate explanations to justify that label. In (Wiegreffe and Marasović, 2021) is given a set of concepts with their definitions and an interesting discussion related to explanations (against different data

sets):

There are three explanations types: *Free-text explanations* - not constrained by rules, *Structured explanations* - built up on constraints injected into the explanation generation process, *Highlights* - defined below; and, based on this classification criterion, there are respectively three types of data sets. A *highlight* is an explanation characterized by the following properties:

Sufficiency, which means that selected explanation words could replace the whole input entry and the model would still generate the same prediction;*compactness* that means short and coherent; *comprehensiveness*, which means that no evidence that supports the prediction is left out; while *plausibility* measures the difference between model-generated highlight and human-generated highlight for the same input and *fidelity* the degree to which the highlight represent decision process of the model.
Unacceptable *plausibility* and other problems, **may have roots beyond the context of AIMLE framework.**

Wiegreffe and Marasović (2021) present a broad and interesting analysis linked to this set of reasons by taking under consideration over 60 datasets and reviewing over 140 research papers. Reasons for not getting accurate outputs from the model can be related to the way annotators act during the collection: annotators' subjectivity; different annotators provide explanations and different ones label or one person can annotate a large data portion instead of a set of annotators used for small slices of data. Because of these, **there is a high risk of not exploiting high diversity of thoughts and instead have biases which are propagated/reflected as artifacts in a large number of predictions.** Authors think that restricting to only one explanation for input, i.e. **not allowing more than one due to doubts for ambiguity incurring during the preparation of data training, may decay model power for generating good explanations and predictions.**
Wiegreffe and Marasović (2021) cite: *"Prasad et al. [98] present a theoretical argument to illustrate that there are multiple ways to highlight input words to explain an annotated sentiment label."*

Giving the model the ability to distinguish and capture different components(aspects, features) in the input text that make up an opinion or image endows the system with additional sharpness to better connect words to corresponding aspects. (McAuley et al., 2012). In fact, authors of (A)IMLE use the BeerAdvocate dataset where user feedback is given separately for each aspect (opinion component), but it is worth knowing the aforementioned argument if it is aimed to increase the applicability of AIMLE.

Multi-Aspect Masker presented in (Antognini et al., 2021) takes under consideration the aforementioned argument; and additionally, tackles the problem of cost increase when the interpretability level gets higher.

Another challenge is the annotators' experience which is not constant nor the same among different people, which translates into rating variance; annotators with a similar level of experience are more likely to provide convergent feedback (McAuley and Leskovec, 2013).

2 Goals and Aims

A key goal is to be a better and more useful person for society by becoming a professional researcher that contributes to humanity by getting prepared to advance my career in academia and industry with the hope of adding some potential to revolutionize the AI field.

One way of achieving this is aiming at being part of a diverse community of scientists that pursue cutting-edge research through the leverage of their expertise, while at the same time, in return investing my time, energy, passion, knowledge, skills, friendliness, and devotion, to contribute to this environment by collaboratively bringing exciting innovations.

Developing and progressing research skills is an obvious key aim.

3 Objectives

Developing state-of-the-art ML/AI techniques is one of the paths to go to achieve the aims.

As shown in the earlier section, gradient estimators(GE) play a crucial role in delivering efficient AI solutions; thus, working for an enhanced GE framework without losing its key advantage, the generality character, is a good research direction.

The objectives are to deeply study and examine how it works with different tasks; analyze the possibility of adding/tunning hyper-parameters; experiment and emerge with framework modifications that increase the performance over competitors. Initially, the experiments and analysis focus on an Annotation/Labeling/Rating system, but the insights gained are relevant not only to it but also to AIMLE which will further get improved thanks to the elaboration of these insights.

"Natural language rationales could provide intuitive, higher-level explanations that are easily understandable by humans, complementing the more broadly studied lower-level explanations based on gradients or attention weights" (Marasović et al., 2020).

As a result of the analysis provided in this paper so far, it becomes clear that this project will lie at the intersections of work on GE and work on RS.

The following list of objectives is imperative to transform aims into reality:

- Keep conducting further research on different GE and RS
- Get additional and advanced training related to (new) concepts that come up during the reading of other's paper
- Enhance AIMLE performance, especially MSE
- Enhance AIMLE to train model simultaneously for different aspects
- Try to increase credibility about the generality success of AIMLE
- Consider continuous latent random variable of exponential family
- Consider other models beyond the exponential family, for instance, those similar but that are not yet expressed via the same formula. **?** lists a few such models
- Try to come up with a closed-form expression that can cover the aforementioned models
- Check whether the generality property for the extended framework still holds
- Publish high-level scientific papers in high-ranked journals and bring positive impact in science, industry, and society
- Eventually, take part in internships

4 Methodology

1. Keep reading from the best journals and keep an eye on the best scientific institutions, inc. research groups of big tech giants, which work on relevant and similar issues to learn more about important and latest findings. Continue to keep updated with weekly magazines such as *The Batch @ DeepLearning.AI of Andrew NG, Towards Data Science, The Sequence of AI knowledge, ScienceDirect, Medical Research Council, etc.*, to learn about AI achievements and emergent needs that AI can help to tackle
2. Use efficiently interdisciplinary training resources and try to Leverage new academic incentives
3. Try and experiment with new approaches and learn new technologies.
4. Initially focus on RS as the first problem where AIMLE framework will be analyzed
5. Select these data sets: BeerAdvocate, E-NSLI (Camburu et al., 2018), E-NSLI-VE (Do et al., 2020) to undertake research and experiments to shed more light on issues related to GE performance and evaluate better its impact on model behavior and correctness of generated explanations by evaluating the plausibility and eventually by studying fidelity.
6. Experiment with models published in GitHub by different authors of the papers
7. Experiment more with AIMLE
8. Jump to the higher points when needed to understand or improve something and hopefully inspire brainstorm
9. To make models using AIMLE train simultaneously for all aspects Multi-Aspect Masker presented in (Antognini at al, 2021a, 2021b) can be considered as one of the baseline examples, although it uses a different approach to the same problem. Compare justification highlights their model generates with the ones AIMLE will produce
10. Consider and address all the issues described in the Annotation/Labeling/Rating Systems section(inc. methodology of data collection) with relevance to our datasets and GE framework.
    - For instance, diversify users/annotators/ppl that give feedback with regards to their social background
    - , For instance, pick up 200-300 training data entries from BeerAdvocate and ascribe their explanations
11. Experiment with different combinations of parameters under different settings.

12. Make the hyper-parameter k related to k-subset distribution a self-adaptive parameter during the model learning
13. Study what other hyper-parameter can be transformed into learning parameters
14. Analyze if NN architecture for the learning to explain experiments (BeerAdvocate) has flaws that propagate further up
15. Using Cloud computing resources for experiments, running on a normal PC is not always a good option as model training takes significant time sometimes. Getting output data faster speeds up the whole analysis process
16. Perform rigorous scientific testing after something changed to assure the rest of the framework is working fine
17. Associate undertaken research with delivering of concept demonstrations
18. Play my role in the research environment with coherence to an attitude that has at its core values such as honesty, patience, supporting one another, and respecting and increasing diversity of thoughts

5 Motivation

Traveling through the World is amazing, but traveling through the Universe is magic. This journey looks impossible unless you pick up the proper navy. When work is combined with studying sciences being part of your heart, satisfaction doubles, and when all of this results in contributing to society, especially in people's health, then satisfaction is not anymore one-sided and gets increased exponentially.

Pursuing research into natural sciences like Math, Machine Learning or any field of AI with a focus on the health/biology realm is about navigating amid a part of mankind; however small that part may be, scientific work may bring pain relief, vanish a pain at all, make life longer and happier and even save more people from death which further means less family mourning and a step more to a healthier and happier society.

Whereas from a professional point of view, however small that part may be, such navigation through it will lead to keenly interesting recognition of some patterns from the Universe. These patterns mean an inter-meshing of interlaced paths, challenges, and excitements. For me, securing the opportunities offered by you is like securing the navy needed for the above navigation; it is also a great privileged manner of serving society.

6 About me

I can say that I have no doubts about my passion for Maths, ML, CNN, NLP, every sub-field of AI, and every field of science. I have a proven will and commitment toward the aims regardless of the hard work they require to be achieved. Regret for selecting the hard work may last at most only several minutes after many hours of work. My so-far relevant training can be seen in my CV. I think that I have a certain familiarity with concepts that such projects involve and that this familiarity gets increased with time, probably due to my will, commitment, and passion which keep me dragged into sciences. I believe that I have the ability to tailor communication skills to different characters and bring to the group additional valuable cooperation and enthusiasm.

7 Ethical issues

Artifacts may contain biases that are reflections of structural features of a given social community with respect to the productions of another community. The same artifacts can be propagated in significant slices of explanations. Diversify the community of annotators/commentators/raters. Every aspect must be at least once associated with an explanation.

8 Timetable

A Gantt chart will be presented due the course.

References

Diego Antognini, Claudiu Musat, and Boi Faltings. Multi-dimensional explanation of target variables from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12507–12515, 2021.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*, 2020.

Justin Domke. Implicit differentiation by perturbation. *Advances in Neural Information Processing Systems*, 23, 2010.

Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. Stochastic optimization of sorting networks via continuous relaxations. *arXiv preprint arXiv:1903.08850*, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Carolyn Kim, Ashish Sabharwal, and Stefano Ermon. Exact sampling with integer linear programs and random perturbations. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Ana Marasović, Chandra Bhagavatula, Jae Sung Park, Ronan Le Bras, Noah A Smith, and Yejin Choi. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. *arXiv preprint arXiv:2010.07526*, 2020.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE, 2012.

Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908, 2013.

Pasquale Minervini, Luca Franceschi, and Mathias Niepert. Adaptive perturbation-based gradient estimation for discrete latent variable models. *arXiv preprint arXiv:2209.04862*, 2022.

Vlad Niculae, Andre Martins, Mathieu Blondel, and Claire Cardie. Sparsemap: Differentiable sparse structured inference. In *International Conference on Machine Learning*, pages 3799–3808. PMLR, 2018.

Mathias Niepert, Pasquale Minervini, and Luca Franceschi. Implicit mle: backpropagating through discrete exponential family distributions. *Advances in Neural Information Processing Systems*, 34:14567–14579, 2021.

George Papandreou and Alan L Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *2011 International Conference on Computer Vision*, pages 193–200. IEEE, 2011.

Max Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, and Chris J Maddison. Gradient estimation with stochastic softmax tricks. *Advances in Neural Information Processing Systems*, 33:5691–5704, 2020.

George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *Advances in Neural Information Processing Systems*, 30, 2017.

Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*, 2021.