Instrumental LLMs

Patrick Power

February 22, 2024

Abstract

In many applied microeconomic contexts, the underlying data is text - think Health Care, Education and Housing. Causal inference in this setting has typically proceeded by hand selecting numerical representations of the text and estimating the corresponding conditional expectation function assuming that treatment or the instrument is locally randomly assigned. Recent developments in Natural Language Processing/AI though have an introduced alternative ways to produce causal estimates from text. In this paper we (1) clarify the general framework for using fine-tuned large language models for causal inference and (2) highlight their relative strengths in the setting of IV with preferential treatment.

1 Introduction

In many applied microeconomic contexts, the underlying data is text - think Health Care, Education, Housing. Causal inference in this setting has typically proceeded by hand selecting numerical representations of the text and estimating the corresponding conditional expectation function assuming that treatment (or instrument) is locally randomly assigned. This vector based approach to casual inference is so ubiquitous that it's often one of the first aspects emphasized in an Econometrics course. Paul Goldsmith-Pinkham writes in Definition Four of Lecture One of his applied PhD Econometrics course that, "We say that D_i is strongly ignorable conditional on a vector X_i."

Recent developments in Natural Language Processing, though, have introduced an alternative way to produce causal estimates from text: fine-tuning Large Language Models. Instead of conditioning on a vector, we can condition on text and estimate the corresponding conditional expectation function via a fine-tuned language model. Figure 1 illustrates this setup in the context of instrumental variables with a binary treatment variable. The instrument (a text document) and the controls (another text document) are the inputs to a language model which is fine-tuned to output the probability of treatment. Together this represents the first stage.

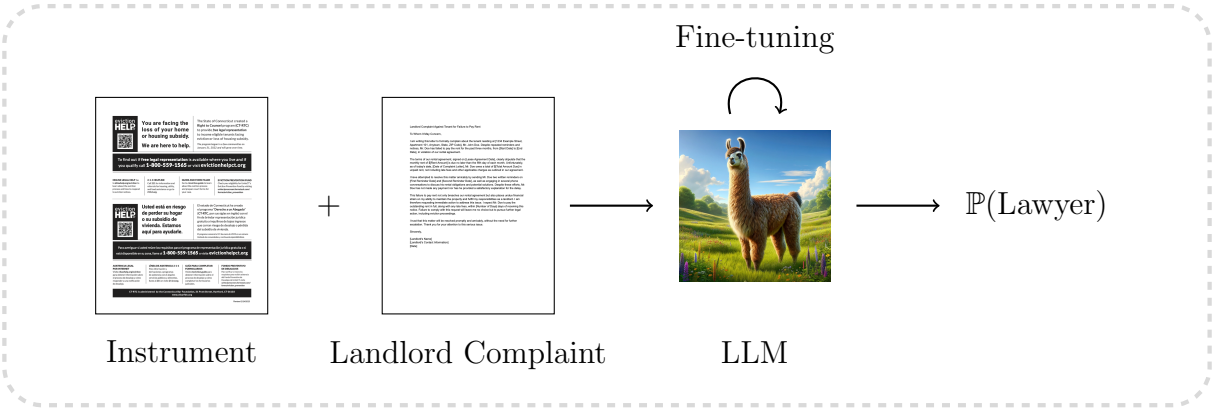


Figure 1: Fine-tuning Large Language Models

In their black-box form, these models do not allow users to reason about a vector representation of the inputs. This can impose a conceptual roadblock for economist who tend to reason(argue!) about the source of the local variation of treatment. As Bengio et al. [2000] notes, these models work over discrete spaces.[1]

In this paper (1) we clarify the conceptual framework for using large language models for causal inference and (2) highlight their relative strengths in the the policy setting where the offer is randomly assigned, and conditional on the offer, the treatment is prioritized. Such settings are are common in practice when treatment is supply constrained as is the case in the rollout of the Right to Counsel across Connecticut, which we'll use as our real

[1]Representing text as a finite sequence of tokens, under the discrete metric this space is separable and complete

world application.

The outline of the paper is as follows. First, we'll introduced large language models and clarify the set of assumptions needed for causal inference. Second, we'll introduce a resulaized instrumental variable framework which allows us to estimate IV parameters using large language models. And third we'll conclude by motivating why large language models might be advantageous relative to linear models.

2 Mostly Harmless Large Language Models

In this section we overview a fundamental challenge of language models and clarify the identification assumptions of causal inference with large language models.

2.1 A Challenge With Large Language Models

The fundamental challenge in language models is that the **product topology** breaks down.[2] This is another way of repeating the common understanding that the context matters which has arguable been best expressed by Firth [1957] who wrote "You shall know a word by the company it keeps." We re-express this notion using the construct of a topology because it the suitable framework for talking about casual identification.

A topology associated with a set is a collection of subsets that satisfies a series of conditions.[3] Given out context, a useful set to equip a topology with is a vocabulary of 50,000 words (or tokens). We can represent the meaning of words by defining an associated topology. For example, one element of the topology may consist of all words which are related to sports. Another element could contain words related to statistics. A topology is the necessary structure in order to convex relationships between words and in particular to highlight which words are similar to others.

Our first exposure to topology likely started with the real numbers. The standard topology associated with this set is defined via the open intervals: $\mathcal{T} = \{(x, y\} \; \forall y \geq x \in \mathcal{R}\}$.[4] One useful property of this topology is that is seamlessly extends to n-tuples of real numbers. That is a point $x := (x_1, x_2)$ is in an element of the topology $\mathcal{B} := (\mathcal{B}_1 \times \mathcal{B}_2)$ if $x_1 \in \mathcal{B}_1$ and $x_2 \in \mathcal{B}_2$. Which is to say that a point is similar to another point if it is similar across each component of the tuple.

Unfortunately, language doesn't cooperate with this product topology. Similar sentences can differ significantly at each word in the sequence. The objective of language models, therefore, is to implicitly learn the a meaningful topology at the sentence level. That is, to map a sequence of dense representation of words (together with the position

[2]Note: The product topology is the coarsest topology under which the projection function, $p_j : \Pi_{i \in J} V_i \to V_j$ is continuous - reference

[3]The set and the empty set are in the topology. The topology is closed under arbitrary unions and under finite intersections.

[4]More precisely, the open intervals are a basis for the standard topology

number) into a dense representation of the sentence. Natural Language Processing papers such as Radford et al. [2018] refer to this as "captur[ing] higher-level semantics."

The focus of this paper, is assuming that this mapping has been learnt, to use a large language model for causal inference.

2.2 Identification

Economists tend to make identification assumptions by arguing about local variation of the treatment. The typical assumption is that across individuals who share similar features, the treatment (or instrument) can be thought of as a good as randomly assigned. That is locally with respect to the controls, there is no selection bias: $\mathbb{E}[\tilde{Y}_i(0)|X_i, D_i = 1] \approx \mathbb{E}[\tilde{Y}_i(0)|X_i, D_i = 0]$. Figure 2 captures this setup where the encoder reflects that the controls are hand selected, identification is based on the local nature of the controls, and a model, typically a linear model, but as flexible as a neural network, is fit to the data to estimate the corresponding Conditional Expectation Function.

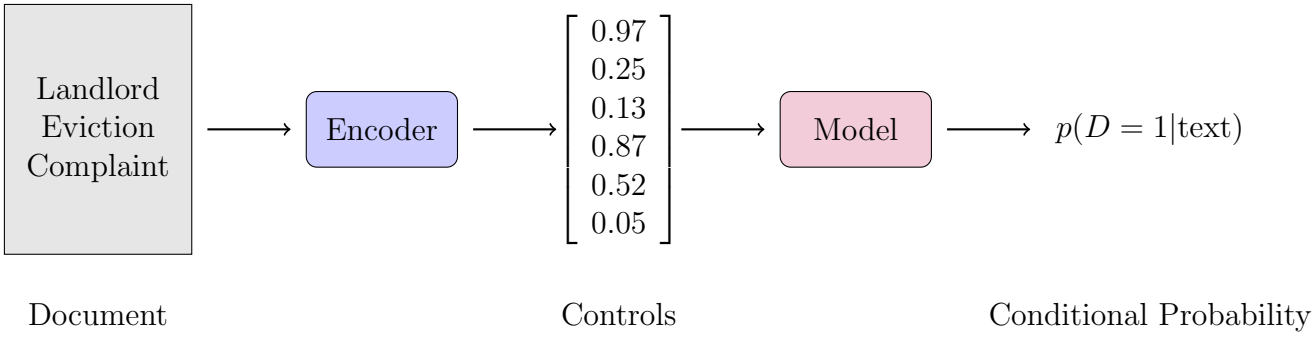
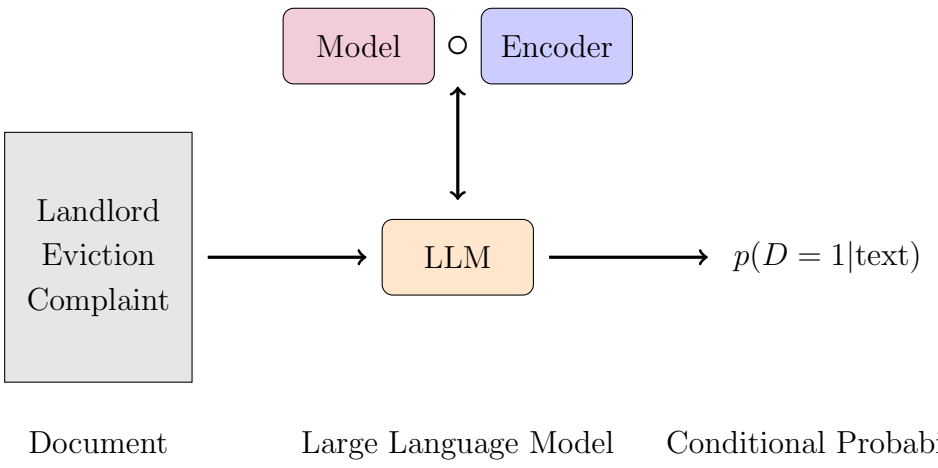


Figure 2: The Standard Pipeline

Black-box large language models don't expose this mapping of words in the sentence into a dense representation of the sentence which means that the notion of similarity can't depend on local euclidean structure.



Below we layout the mathematical definition and assumption for causal identification. We begin with the underlying probability space which is a tuple of the sample set (a set containing all possible units of observation), a σ-algebra, and the population probability measure.

$$\big(\Omega,\, \mathcal{F}_{\Omega}, \mathbb{P}\big)$$

On this space, we define the following three random variables of interest: Controls, Treatment, and Potential Outcomes, where for simplicity we assume that both treatment and outcomes are binary.

$$\begin{aligned} X_i &: \Omega \to \mathrm{L}^{\mathrm{V}} \\ D_i &: \Omega \to \{0,1\} \\ \tilde{Y}_i &: \{0,1\} \to \Omega \to \{0,1\} \end{aligned}$$

The key aspect is that the random variable X transforms the underlying probability space into a probability space defined over text (finite sequence of tokens of length L from a vocabulary of V tokens).

$$\big(\Omega, \mathcal{F}_{\Omega},\, \mathbb{P}\big) \xmapsto{X} \big(\mathrm{L}^{\mathrm{V}}, \mathcal{F}_{\mathrm{L}^{\mathrm{V}}}, \mathbb{P} \circ X^{-1}\big)$$

In addition to defining X, though, the researcher must implicitly define$\mathcal{F}_{\mathrm{L}^{\mathrm{V}}}$ which captures the semantic meaning of sentences. Although it's possible given that L^{V}, we don't argue for taking the discrete topology. Then with this construct, the identification assumption is that treatment is conditionally independent of the potential outcomes with respect to $\mathcal{F}_{\mathrm{L}^{\mathrm{V}}}$. That is,

$$\forall A, B, C \in \sigma(\tilde{Y}_i), \sigma(D_i), \sigma(X_i), \quad \mathbb{P}_C(A \cap B) = \mathbb{P}_C(A)\mathbb{P}_C(B)$$

Under this conditional independence assumption, the corresponding conditional expectation function has a casual interpretation

$$\text{Average Treatment Effect } = \int \mathbb{E}[Y_i | D_i = 1, X_i] - \mathbb{E}[Y_i | D_i = 0, X_i] d\mathbb{P}$$

3 Framework

Having clarified the formal framework for causal inference with text based models, we'll restrict our focus for the rest of the paper to the residualizded approach to causal inference and in particular residualized instrumental variables.

3.1 Residualized Models

The typical approach is two stage least square where we first predict treatment using a linear model of the controls X and the instrument Z. And then in the second stage, fit a least squares model where the treatment variable has been replaced by its fitted counterpart.

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + \beta_2 X_i + \varepsilon_i, \quad \hat{D}_i = \hat{\gamma}_1 X_i + \hat{\gamma}_z Z_i \tag{1}$$

Under the Frish Waugh Lovell Theorem this two step procedure can be understood as regressing the outcome variable on a single residualized variable wherer $\bar{\hat{D}}_i$ is the fitted value of the first stage predictions based only on the controls.

$$Y_i = \beta_1 \big(\hat{D}_i - \bar{\hat{D}}_i\big) + \eta_i \tag{2}$$

We can introduce a non-parametric equivalent of the same setup by replacing the linear models with the associated conditional expectation functions. This residualized term highlights the essence of the instrumental variable strategy: we use only the local variation of the treatment variable which is generated by the instrument.

$$Y_i = \beta_1 \big(\mathbb{E}[D_i|X_i, Z_i] - \mathbb{E}[D_i|X_i]\big) + \varepsilon_i \tag{3}$$

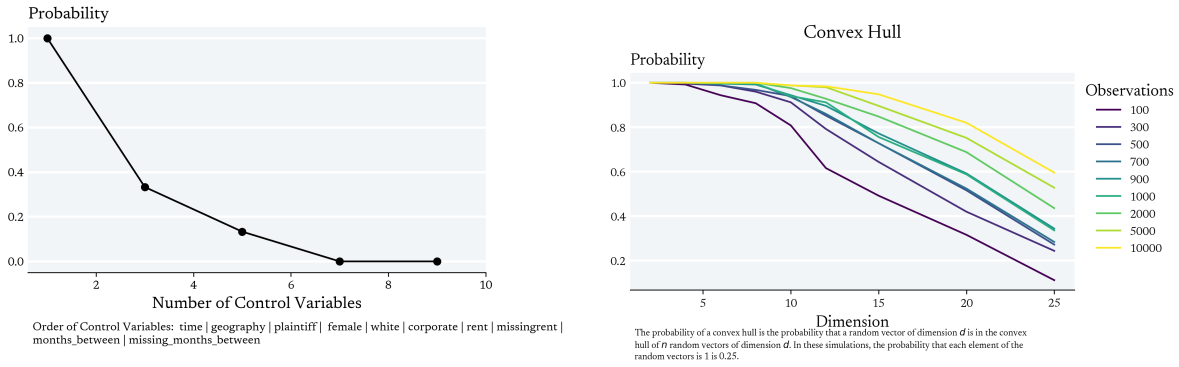
4 Motivation

In the finite sample, the success of these language models (as with all models) depends on their ability generalize (Balestriero et al. [2021]). Despite having within variation of the treatment at the population level, the need to generalize in the finite sample is clear – in a "typical" empirical settings, observations from the treated group lie outside the convex hull formed from observations in the control group (figure 3).

One motivation for linear models then is that as kernel methods, we understand how they generalize because we've specified the similarity function between observations (the inner product). Conversely neural networks, and large language models are black box because the similarity function is learnt during the training phase of the model (Domingos [2020]): $K(x, x') = \int_{c(t)} \nabla_\theta f_\theta(x)^T \nabla_\theta f_\theta(x') dt$.

One reason for not relying on linear models is that they can generalize poorly. As in the case of residualized IV (explained above), those with an instrument = 1 should have positive residuals while those with the instrument set to 0 should have negative residuals. As figure 4a illustrates, the linear model, by underfitting the data, doesn't necessarily satisfy this condition. Several individuals with the instrument set to 0 have positive residuals and vice-versus.[5] The non-linear language model, meanwhile has no problem satisfying the

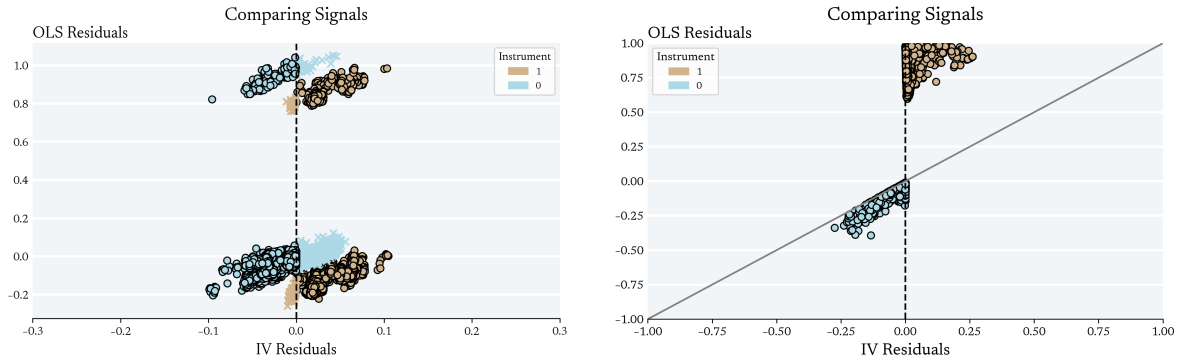
[5] Using data from The Right to Counsel at Scale



(a) Real Data (b) Simulated Data

Figure 3: (Left) With a sample size greater than $10,000$ observations, with just a small number of controls, the probability that an observation is within the convex hull of the other observations diminishes significantly. (Right) Using simulated data, but focused on dimensions rather then control variables, the same pattern occurs.

monotonicity assumption. It cleanly exploits the meaningful variation generated by the instrument.



(a) Linear Model (b) Non-linear Model

Figure 4: Linear models underfit the data which can lead them to get the wrong sign of the residuals. Nonlinear models don't tend to underfit the data but of course are more likely to overfit the data (not shown).

In certain policy contexts we may have information about who is likely to take up the treatment. This is most likely in cases where treatment is a result of both an individual accepting the treatment, and the offer of the treatment being prioritized due to limited supply. For example, consider the context of the rollout of the Right to Counsel across Connecticut. Free legal representation was available in only certain zip codes (the instrument). However within these zip codes, given the limited supply of legal aid relative to the demand, tenants with vouchers and disabilities were often prioritized over other tenants (based on conversations with legal aid lawyers).

Language models can leverage this information to improve the precision of the first stage estimate. Mechanically we can do so by prepending each text observation with a description of how treatment is prioritized. As Figure 5 highlights fine-tuned language

models with prompting can take advantage of this information – outperforming a standard large language model when the first stage noise is low – and yet is flexible enough to ignore this knowledge if it turns out not to be true – outperforming a fixed large language model when the first stage noise is high.

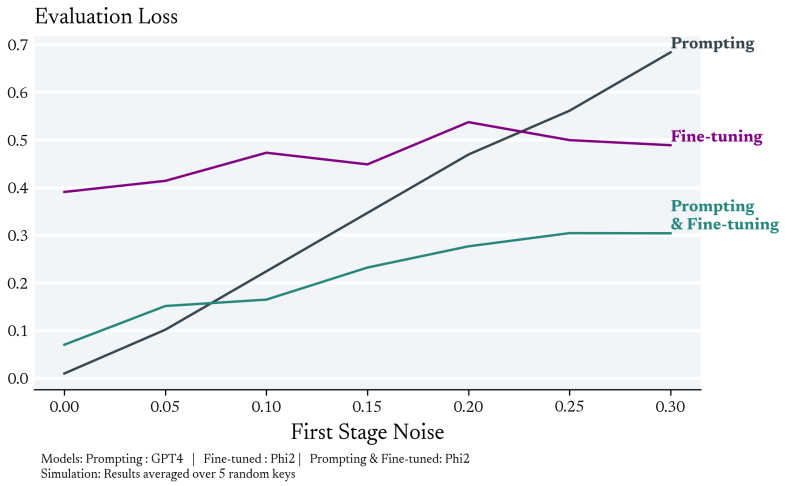


Figure 5: Optimal Evaluation Loss with Early Stopping

5 Conclusion

In many contexts in applied microeconomics, conceptually, the most suitable set of controls is a textual document. One could imagine that if we were studying evictions and had access to the landlord's complaint against the tenant or were studying the Low Income Housing Tax Credit and were privy to the developers application, that if the treatment of interest was randomly assigned between semantically similar documents, we would have a great deal of confidence in our estimates.

Large Language models make this type of casal identification strategy feasible. In this paper we (1) clarify the conceptual framework of causal identification with text based models and (2) highlight their relative advantage when treatment is prioritized.

Summary

For identification, treatment should be conditionally independent of the potential outcomes with respect of a semantically meaningful σ-algebra defined on the space of text.

In the case of IV where the instrument is randomly assigned but the treatment is prioritized (due to limited supply for example), large language models can more precisely estimate the first stage by incorporating the prioritization reasons in the prompt.

References

Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*, 2021.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Vicki Boykis. What are embeddings, 2024. URL https://vickiboykis.com/what_are_embeddings/. Accessed: 2024-01-07.

Pedro Domingos. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

John Rupert Firth. Ethnographic analysis and language with reference to malinowski's views. *Man and Culture: an evaluation of the work of Bronislaw Malinowski*, pages 93–118, 1957.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Graham Neubig. Introduction to NLP. YouTube, 2024.

Omar Osanseviero. Understanding sentence embeddings. https://osanseviero.github.io/hackerllama/blog/posts/sentence_embeddings/, 2024. Accessed: 02-08-2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V Le. Inverse scaling can become u-shaped. *arXiv preprint arXiv:2211.02011*, 2022a.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022b.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L Griffiths. Recovering mental representations from large language models with markov chain monte carlo. *arXiv preprint arXiv:2401.16657*, 2024.

Bruno Zimmermann. Ictp diploma - topology. YouTube video, 2016. URL https://www.youtube.com/watch?v=28BluiBRdUk&t=1599s. Featured playlist of 20 videos. ICTP Mathematics.

6 Appendix

$$\begin{aligned}
\text{Residual}(Z=1,X) &= \mathbb{E}[D|Z=1,X] - \mathbb{E}[D|X] \\
&= \mathbb{E}[D|Z=1,X] - \big(\mathbb{P}(Z=1|X)\mathbb{E}[D|Z=1,X] \\
&\quad + (1-\mathbb{P}(Z=1|X)\mathbb{E}[D|Z=0,X]\big) \\
&= (1-\mathbb{P}(Z|X))\big(\mathbb{E}[D|Z=1,X] - \mathbb{E}[D|Z=0,X]\big) \\
&= \underbrace{(1-\mathbb{P}(Z|X))\big(\mathbb{P}(\text{Complier}|X)}_{\geq 0}
\end{aligned}$$

$$\text{Residual}(Z=0,X) = \underbrace{-\mathbb{P}(Z=1|X)\big(\mathbb{P}(\text{Complier}|X)}_{\leq 0}$$