

- Mitchell, E., Lin, C., Bosselut, A., Finn, C., and Manning, C. D. Fast model editing at scale. In *International Conference on Learning Representations*, 2021.
- Pearl, J. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 411–420, 2001.
- Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.
- Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL <https://aclanthology.org/D19-1250>.
- Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H., and Riedel, S. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*, 2020.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, pp. 9, 2019.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- Roberts, A., Raffel, C., and Shazeer, N. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL <https://aclanthology.org/2020.emnlp-main.437>.
- Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020a.
- Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. M. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020b.
- Wang, B. and Komatsuzaki, A. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., and Dolan, W. B. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS*, 2018.
- Zhao, S., Pascual, D., Brunner, G., and Wattenhofer, R. Of non-linearity and commutativity in BERT. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.
- Zhong, Z., Friedman, D., and Chen, D. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5017–5033, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.398. URL <https://aclanthology.org/2021.naacl-main.398>.
- Zhu, C., Rawat, A. S., Zaheer, M., Bhojanapalli, S., Li, D., Yu, F., and Kumar, S. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*, 2020.

# Appendices

## A Solving for $\Lambda$ Algebraically

Here we present the detailed derivation of Eqn. 2, including the linear system that is used to calculate  $\Lambda$  from  $v_*$ ,  $C$ , and  $k_*$ . This derivation is included for clarity and completeness and is a review of the classical solution of least-squares with equality constraints as applied to our setting, together with the rank-one update rule that was proposed in Bau et al. (2020).

We assume that  $W$  is the optimal least-squares solution for memorizing a mapping from a previous set of keys  $K$  to values  $V$ ; this solution can be written using the normal equations as follows.

$$\text{the } W \text{ that minimizes } \|WK - V\|_F^2 \quad (5)$$

$$\text{solves } WKK^T = VK^T \quad (6)$$

Here the Frobenius norm is used to write the total square error since the variable being optimized is a matrix  $W$  rather than a vector  $x$  as in the classical textbook presentation of least squares.

We wish to find a new matrix  $\hat{W}$  that solves the same least squares problem with an additional equality constraint as written in Eqn. 2:

$$\hat{W}k_* = v_* \quad (7)$$

This is the well-studied problem of least squares with a linear equality constraint. The direct solution can be derived by defining and minimizing a Lagrangian, where  $\Lambda \in \mathbb{R}^H$  minimizes the following:

$$\text{define } L(\hat{W}, \Lambda) = \frac{1}{2}\|\hat{W}K - V\|_F^2 - \Lambda^T(\hat{W}k_* - v_*) \quad (8)$$

$$= \frac{1}{2}(\hat{W}K)(\hat{W}K)^T - V(\hat{W}K)^T + \frac{1}{2}VV^T - \Lambda^T(\hat{W}k_* - v_*) \quad (9)$$

$$\text{setting } 0 = \frac{\partial L}{\partial \hat{W}} = \hat{W}(KK^T) - VK^T - \Lambda k_*^T \quad (10)$$

$$\hat{W}KK^T = VK^T + \Lambda k_*^T \quad (11)$$

Subtracting Eqn. 6 from Eqn. 11, most terms cancel, and we obtain the update rule:

$$(\hat{W} - W)KK^T = \Lambda k_*^T \quad (12)$$

$$\hat{W} = W + \Lambda(C^{-1}k_*)^T \quad (13)$$

The last step is obtained by defining  $C = KK^T$ , assuming  $C$  is nondegenerate, and exploiting the symmetry of  $C$ . Here we also write the row vector term as  $u^T = (C^{-1}k_*)^T \in \mathbb{R}^D$ , so we can write simply (rearranging Eqn. 2 and Eqn. 13):

$$\hat{W}I - \Lambda u^T = W \quad (14)$$

To solve for  $\Lambda$ , we note that Eqn. 14 and Eqn. 7 form a linear system that allows both  $\hat{W}$  and  $\Lambda$  to be solved simultaneously if written together in block form.

$$\left[ \begin{array}{c|c} \hat{W} & \Lambda \end{array} \right] \left[ \begin{array}{c|c} I & k_* \\ \hline -u^T & 0 \end{array} \right] = \left[ \begin{array}{c|c} W & v_* \end{array} \right] \quad (15)$$

That is equivalent to substituting Eqn. 13 into Eqn. 7 and calculating the following:

$$\hat{W}k_* = (W + \Lambda u^T)k_* = Wk_* + \Lambda(u^T k_*) = v_* \quad (16)$$

$$\Lambda = \frac{v_* - Wk_*}{u^T k_*} = \frac{v_* - Wk_*}{(C^{-1}k_*)^T k_*} \quad (17)$$

## B Causal Tracing

### B.1 Experimental Settings

Note that, in by-layer experimental results, layers are numbered from 0 to  $L - 1$  rather than 1 to  $L$ .

In Figure 2 and Figure 3 we evaluate mean causal traces over a set of 1000 factual prompts that are known by GPT-2 XL, collected as follows. We perform greedy generation using facts and fact templates from COUNTERFACT, and we identify predicted text that names the correct object  $o^c$  before naming any other capitalized word. We use the text up to but not including the object  $o^c$  as the prompt, and we randomly sample 1000 of these texts. In this sample of known facts, the predicted probability of the correct object token calculated by GPT-2 XL averages 27.0%.

In the corrupted run, we corrupt the embeddings of the token naming the subject  $s$  by adding Gaussian noise  $\epsilon \sim \mathcal{N}(0; \nu)$ , where  $\nu = 3\sigma_t$  is set to be three times larger than the observed standard deviation  $\sigma_t$  of token embeddings as sampled over a body of text. For each run of text, the process is repeated ten times with different samples of corruption noise. On average, this reduces the correct object token score to 8.47%, less than one third the original score.

When we restore hidden states from the original run, we substitute the originally calculated values from the same layer and the same token, and then we allow subsequent calculations to proceed without further intervention. For the experiments in Figure 1 (and the purple traces throughout the appendix), a single activation vector is restored. Naturally, restoring the last vector on the last token will fully restore the original predicted scores, but our plotted results show that there are also earlier activation vectors at a second location that also have a strong causal effect: the average maximum score seen by restoring the most impactful activation vector at the last token of the subject is 19.5%. In Figure 1j where effects are bucketed by layer, the maximum effect is seen around the 15th layer of the last subject token, where the score is raised on average to 15.0%.

### B.2 Separating MLP and Attn Effects

When decomposing the effects into MLP and Attn lookups, we found that restoring single activation vectors from individual MLP and individual Attn lookups had generally negligible effects, suggesting the decisive information is accumulated across layers. Therefore for MLP and Attn lookups, we restored runs of ten values of  $m_i^{(l)}$  (and  $a_i^{(l)}$ , respectively) for an interval of layers ranging from  $[l_* - 4, \dots, l_* + 5]$  (clipping at the edges), where the results are plotted at layer  $l_*$ . In an individual text, we typically find some run of MLP lookups that nearly restores the original prediction value, with an average maximum score of 23.6%. Figure 2b buckets averages for each token-location pair, and finds the maximum effect at an interval at the last entity token, centered at the 17th layer, which restores scores to an average of 15.0%. For Attn lookups (Figure 2c), the average maximum score over any location is 19.4%, and when bucketed by location, the maximum effect is centered at the 32nd layer at the last word before prediction, which restores scores to an average of 16.5%.

Figure 7 shows mean causal traces as line plots with 95% confidence intervals, instead of heatmaps.

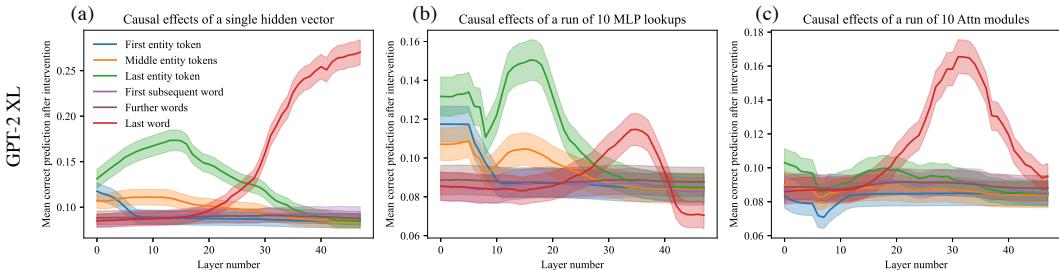


Figure 7: Mean causal traces of GPT-XL over a sample of 1000 factual statements, shown as a line plot with 95% confidence intervals. (a) Shows the same data as Figure 1j as a line plot instead of a heatmap; (b) matches Figure 1k; (c) matches Figure 1m. The confidence intervals confirm that the distinctions between peak and non-peak causal effects at both early and late sites are significant.