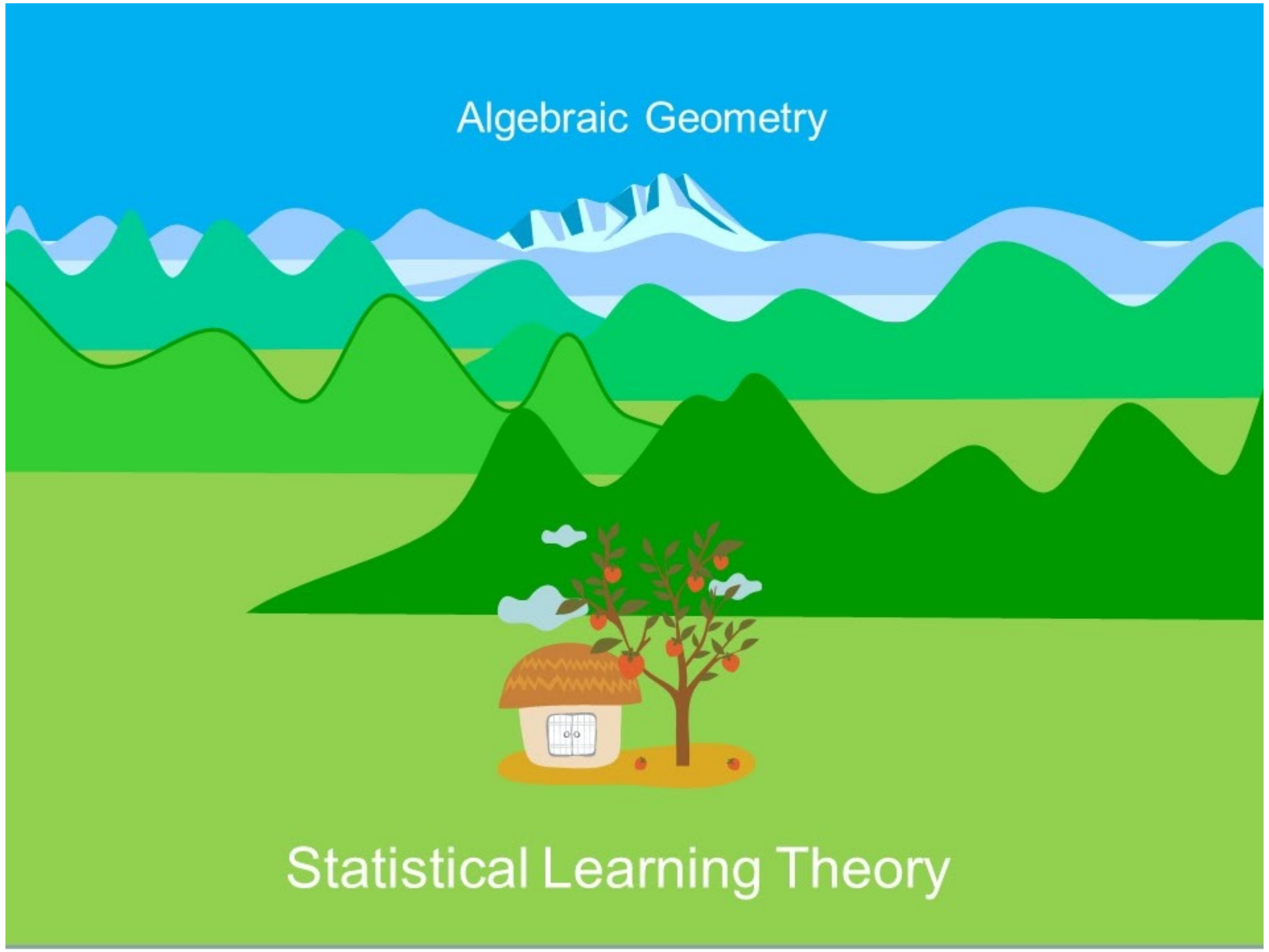
SLT for Alignment

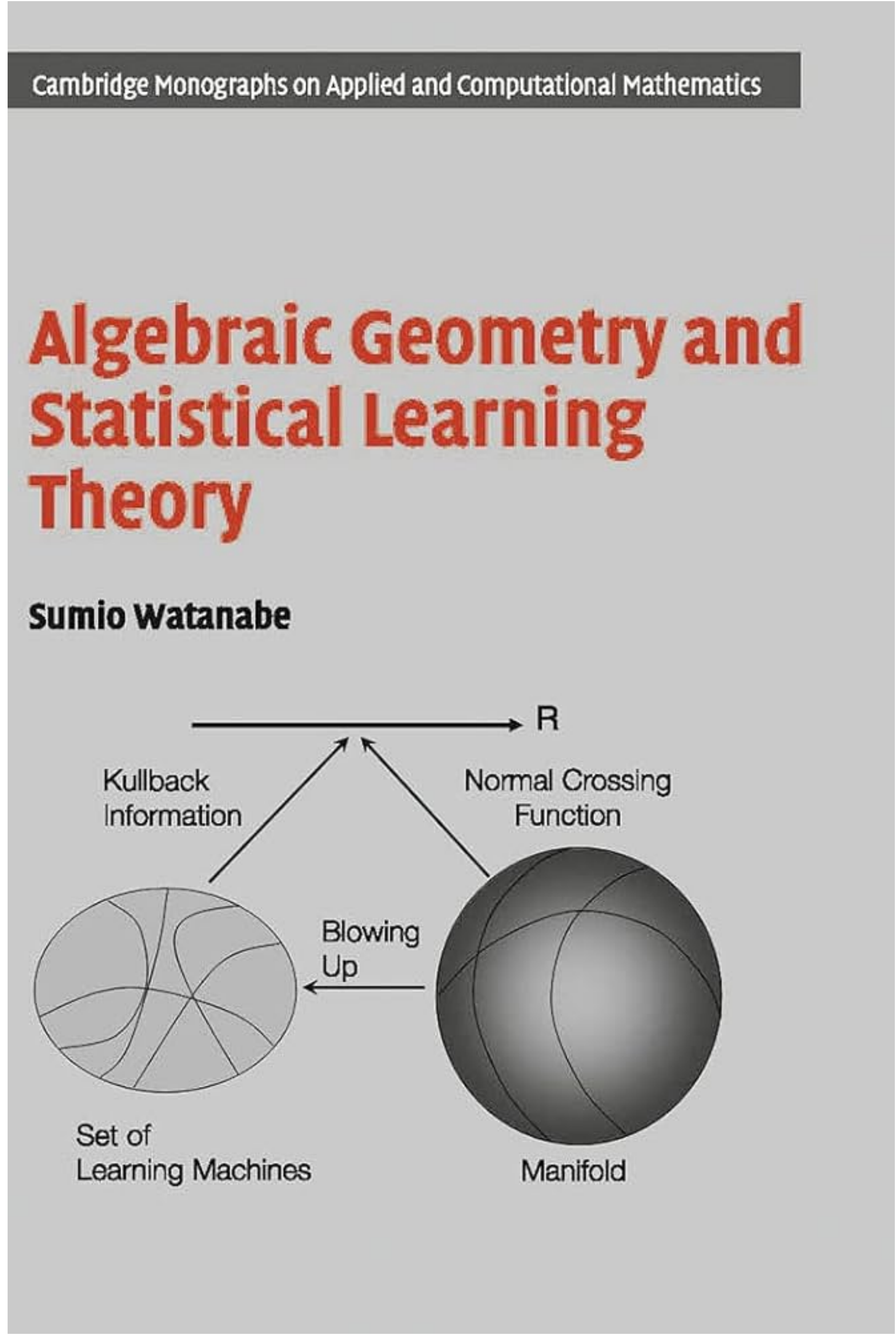
Theoretical and Empirical Aspects

Daniel Murfet
18/4/2025
Timaeus

SLT = Singular Learning Theory

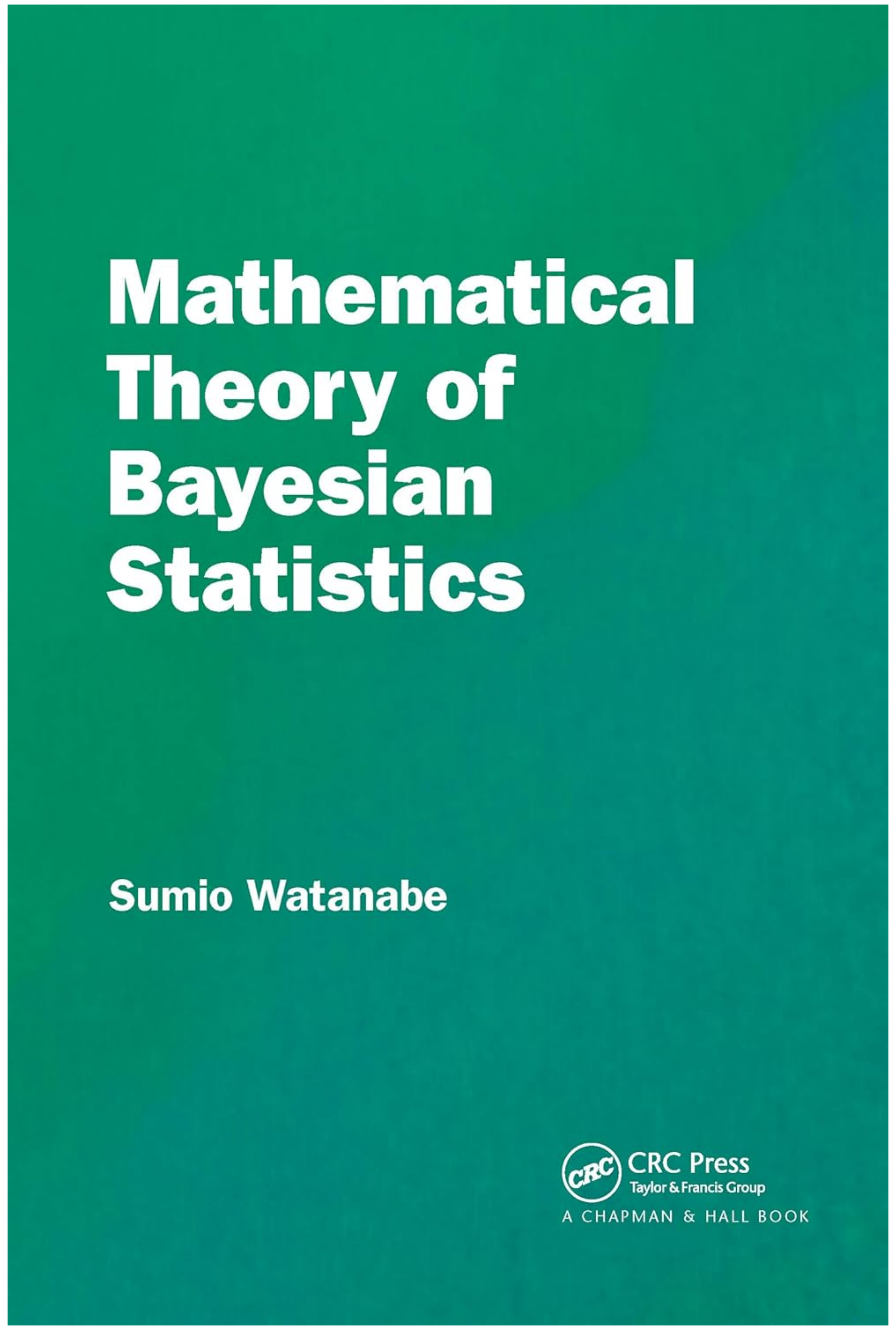
Sumio Watanabe




Cambridge Monographs on Applied and Computational Mathematics

Algebraic Geometry and Statistical Learning Theory

Sumio Watanabe


Mathematical Theory of Bayesian Statistics

Sumio Watanabe

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

https://sites.google.com/view/sumiowatanabe/home

Alignment

- The capabilities and alignment of modern AI models are **specified indirectly**, by engineering the data distribution used to train them
- However, model behaviour is **underspecified**: many different ways of performing the computation within the model are consistent with low loss
- Two models with the same evaluation performance but different “modes of computation” may **behave differently** in other settings (e.g. reward hack or not)
- Many approaches to alignment benefit from progress on **understanding these differences**, detecting them and tracing them back to causes
- SLT has non-trivial things to say here, and empirical tools to back it up

“You are what you eat: AI alignment requires understanding how data shapes structure and generalisation” S. Pepin Lehalleur, J. Hoogland, M. Farrugia-Roberts, S. Wei, A. Gietelink Oldenziel, G. Wang, L. Carroll, D. Murfet

1. Model Selection
2. Some Theory
3. Experiment
4. Structure

1. Model Selection

Let D_n be a dataset of size n and let $L_n(w)$ be the empirical loss for a statistical model $p(x \mid w)$ parametrised by w with prior $\varphi(w)$

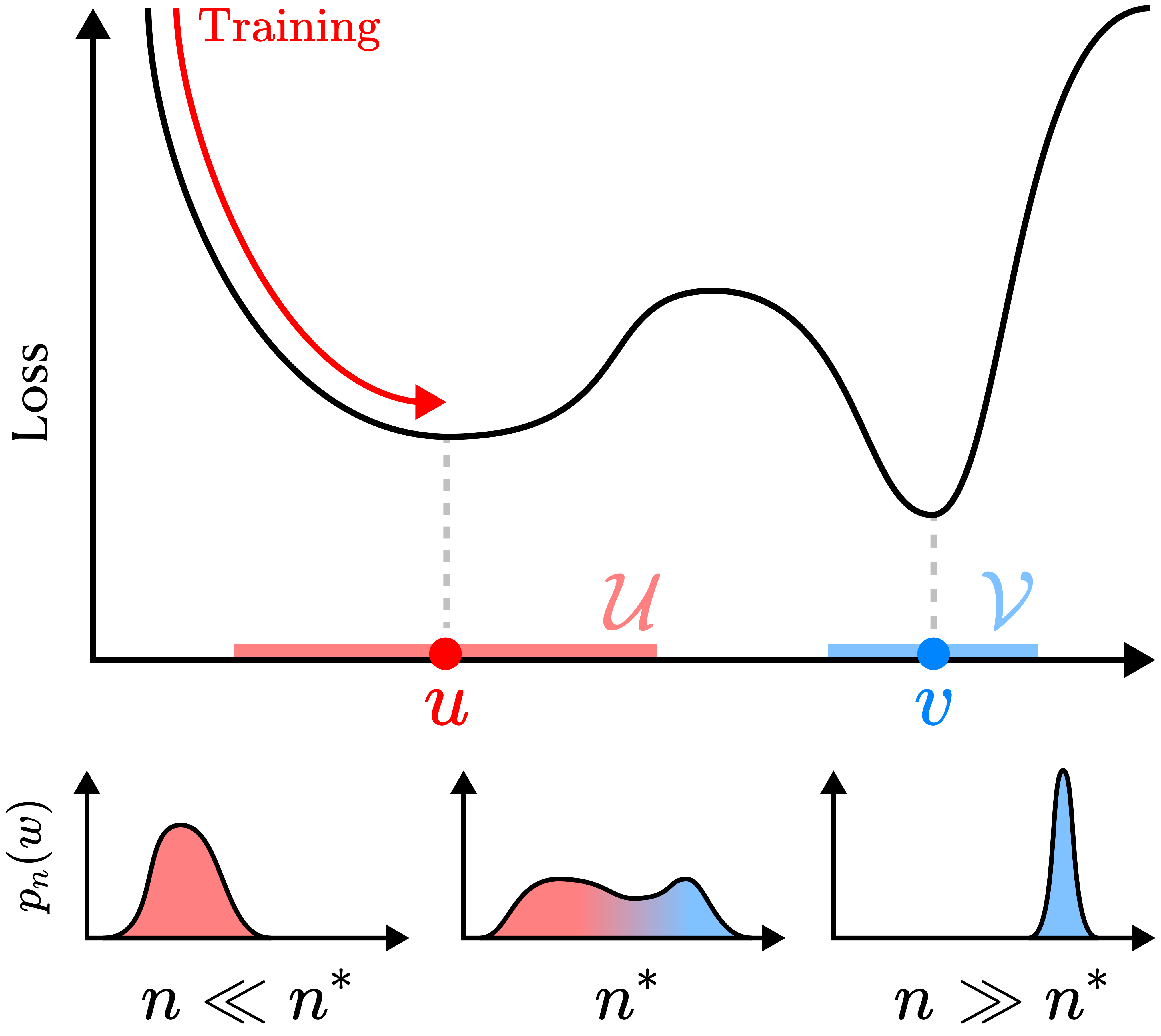
$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i \mid w)$$

$$p(w \mid D_n) = \frac{1}{Z_n}\prod_{i=1}^{n} p(X_i \mid w)\varphi(w)$$

$$p(w \in \mathscr{U} \mid D_n) = \int_{\mathscr{U}} p(w \mid D_n) dw$$

$$= \frac{1}{Z_n} \underbrace{\int_{\mathscr{U}} \prod_{i=1}^{n} p(X_i \mid w)\varphi(w)}_{\exp(-F_n(\mathscr{U}))}$$

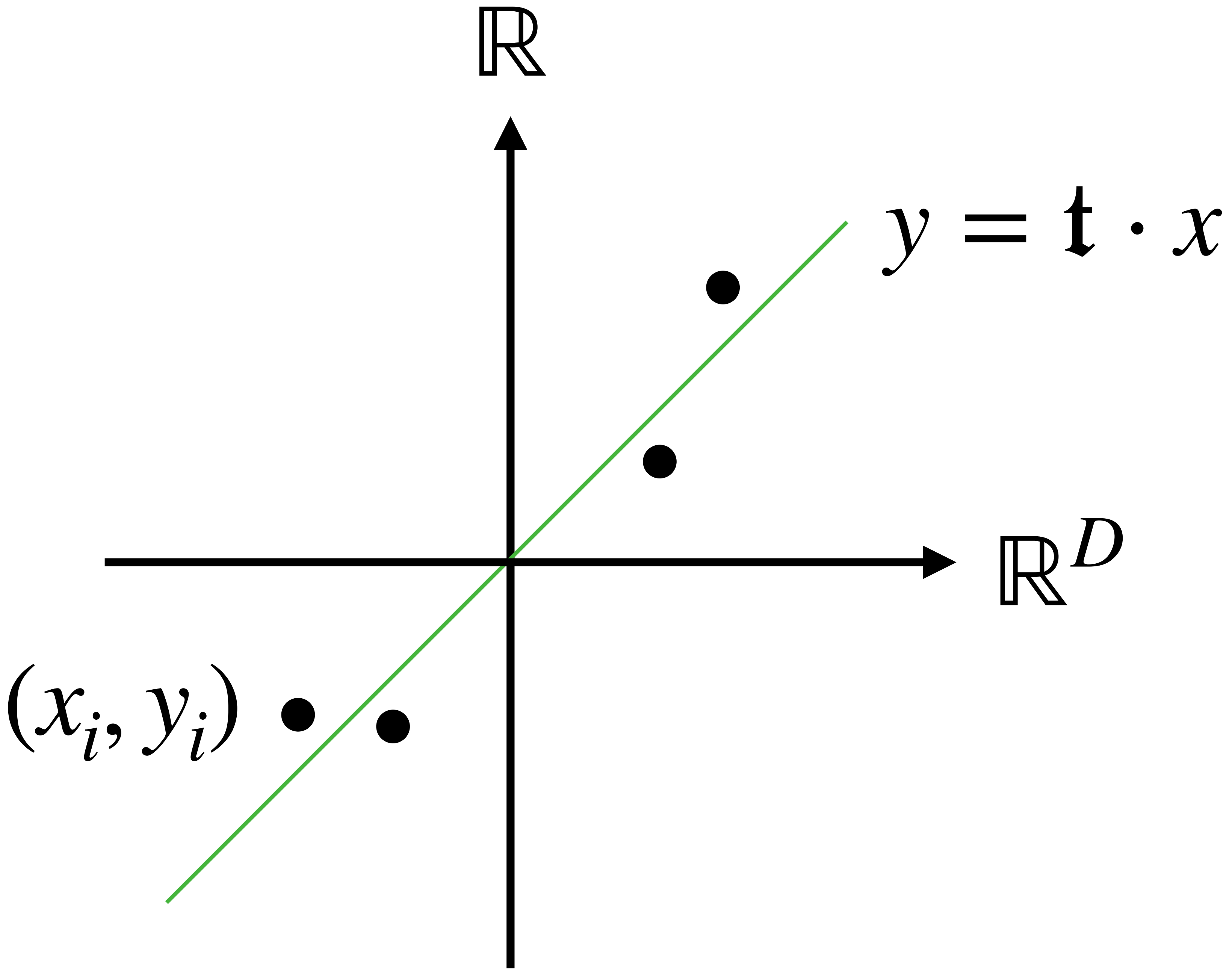
Free energy

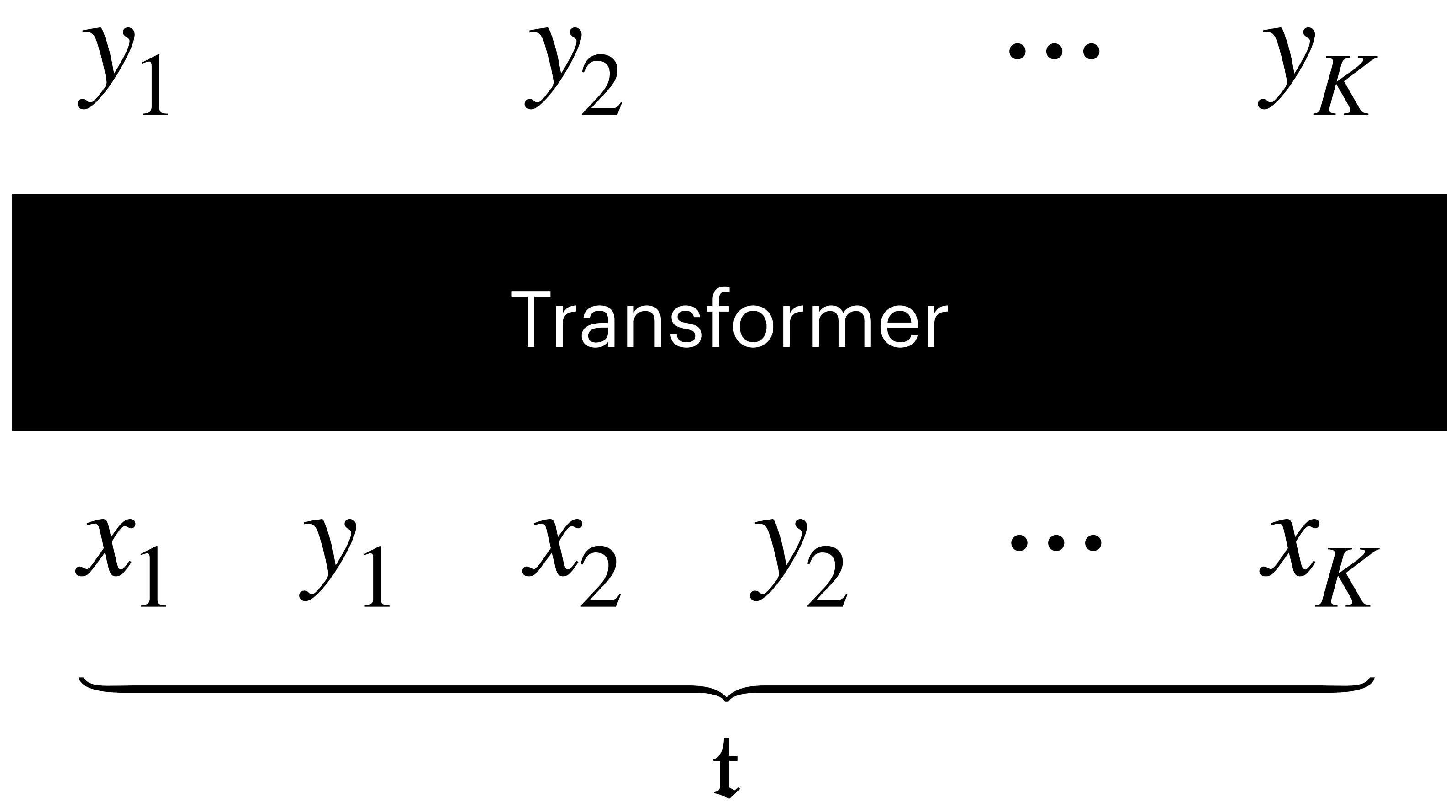


$$F_n(\mathscr{U}) := -\log \int_{\mathscr{U}} \prod_{i=1}^{n} p(X_i \mid w)\varphi(w)dw$$

$$\log \frac{p(w \in \mathscr{V} \mid D_n)}{p(w \in \mathscr{U} \mid D_n)} = F_n(\mathscr{U}) - F_n(\mathscr{V})$$

But does this actually happen?

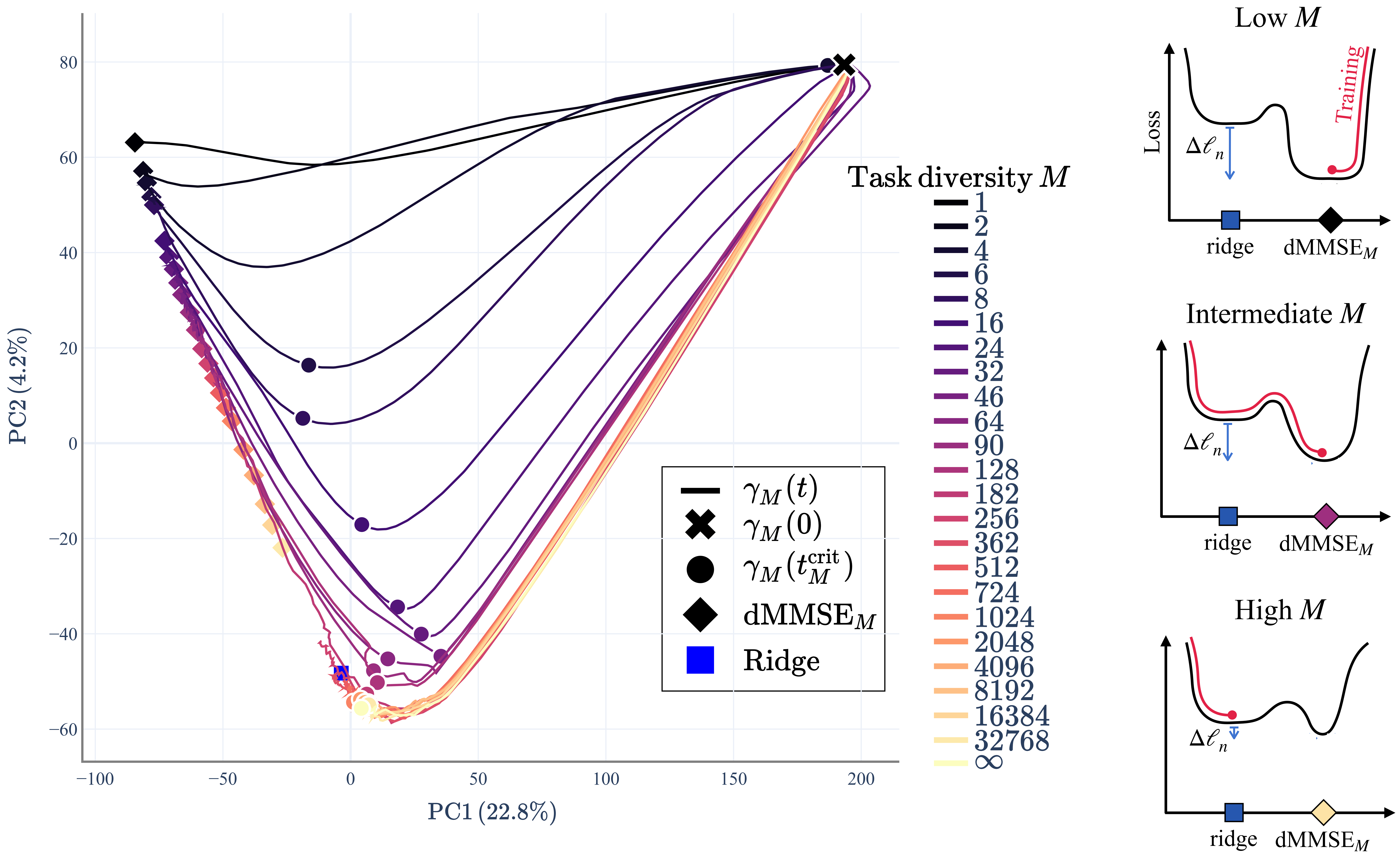


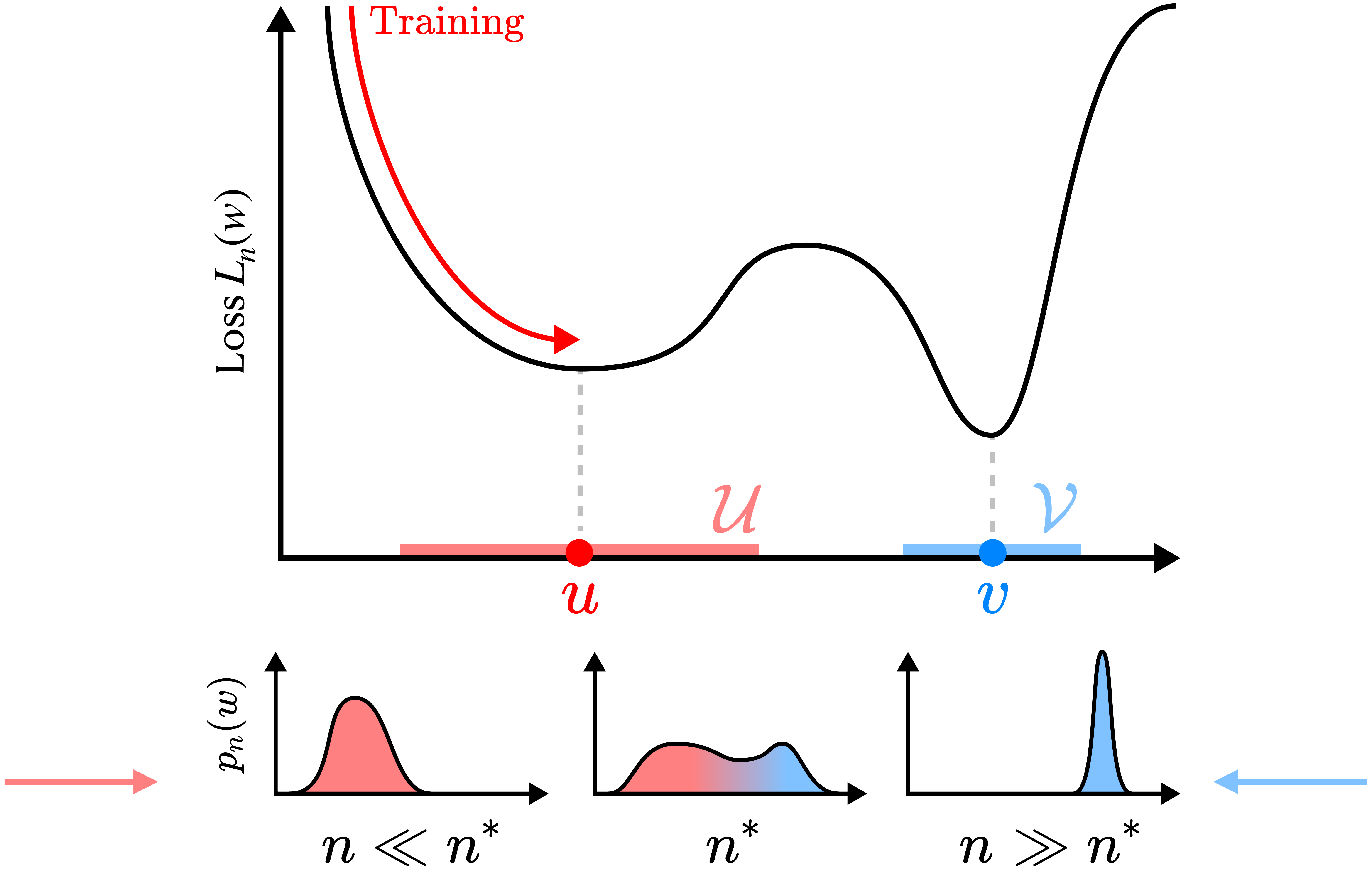

y_1 y_2 $\cdots$ y_K
Transformer
x_1 y_1 x_2 y_2 $\cdots$ x_K
$\mathfrak{t}$

$$y_i = \mathfrak{t} \cdot x_i + \epsilon$$

- Following Garg et al 2022 and Raventós et al 2023, we study in-context linear regression
- A transformer is trained to predict $y \in \mathbb{R}$ given $x \in \mathbb{R}^D = \mathbb{R}^8$ with *task diversity* M
- dMMSE$_M$ = posterior predictive distribution on $\{\mathfrak{t}_1, \ldots, \mathfrak{t}_M\}$ (memorising solution)
- Ridge = posterior predictive distribution on W for $\mathfrak{t} \sim N(0, I_D)$ (generalising solution)

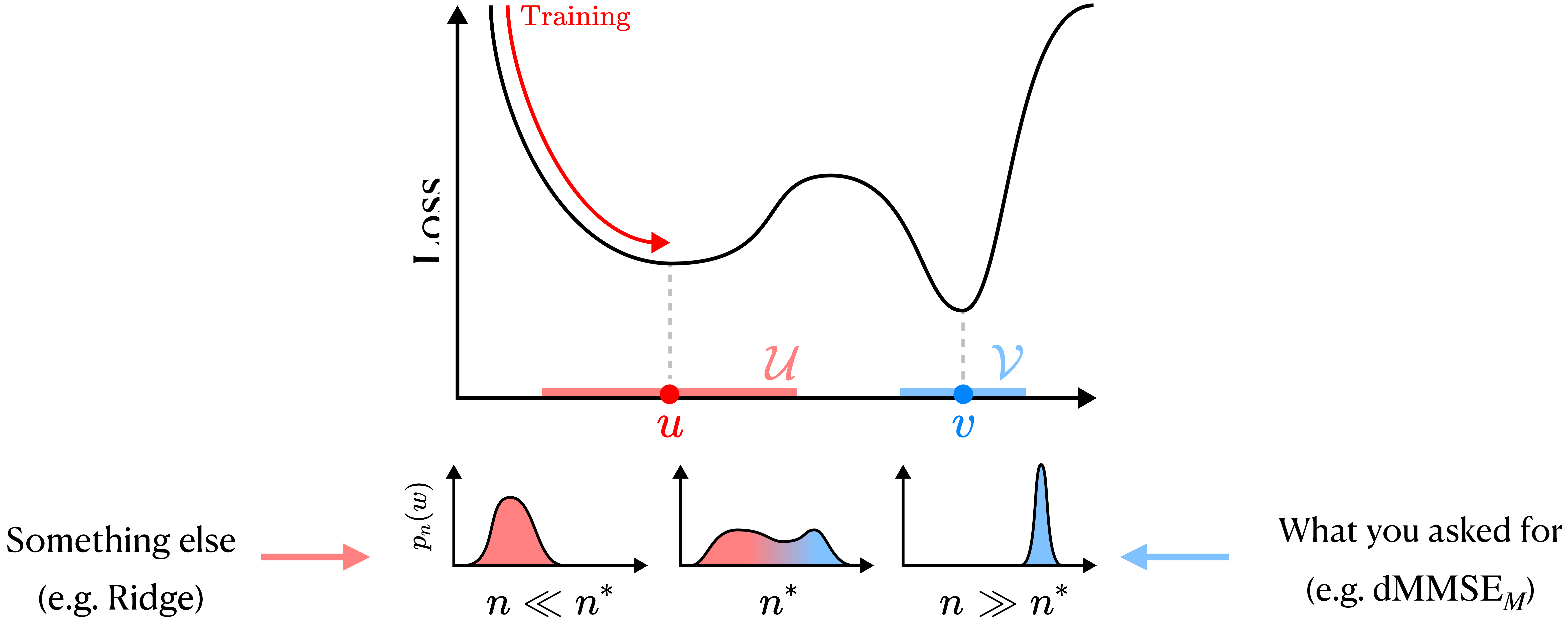
"Dynamics of transient structure in in-context linear regression transformers" L. Carroll, J. Hoogland, M. Farrugia-Roberts, D. Murfet 2025


PC2 (4.2%)
80
60
40
20
0
−20
−40
−60
PC1 (22.8%)
−100
−50
0
50
100
150
200
$\gamma_M(t)$
$\gamma_M(0)$
$\gamma_M(t_M^{\text{crit}})$
dMMSE$_M$
Ridge
Task diversity M
1
2
4
6
8
16
24
32
46
64
90
128
182
256
362
512
724
1024
2048
4096
8192
16384
32768
∞
Low M
Loss
Training
$\Delta\ell_n$
ridge
dMMSE$_M$
Intermediate M
$\Delta\ell_n$
ridge
dMMSE$_M$
High M
$\Delta\ell_n$
ridge
dMMSE$_M$



Something else
(e.g. Ridge)

What you asked for
(e.g. dMMSE$_M$)



Something else
(e.g. Ridge)

What you asked for
(e.g. dMMSE$_M$)

- The *optimal* solution (in a Bayesian sense) is not necessarily the *lowest loss* solution
- For feasible *n* your desired solution could be “screened” by lower complexity solutions
- In a paradigm where you indirectly specify safe behaviour by engineering the data distribution, the nature of this screening seems important
- Is reward hacking lower complexity?

2. Some Theory

Defining the Learning Coefficient

- Let $L(w) = -\int q(x) \log p(x \mid w)$ be the average negative log likelihood (pop. loss)
- The *local learning coefficient* (LLC) at a local minimum $w^* \in W$ of $L(w)$ is

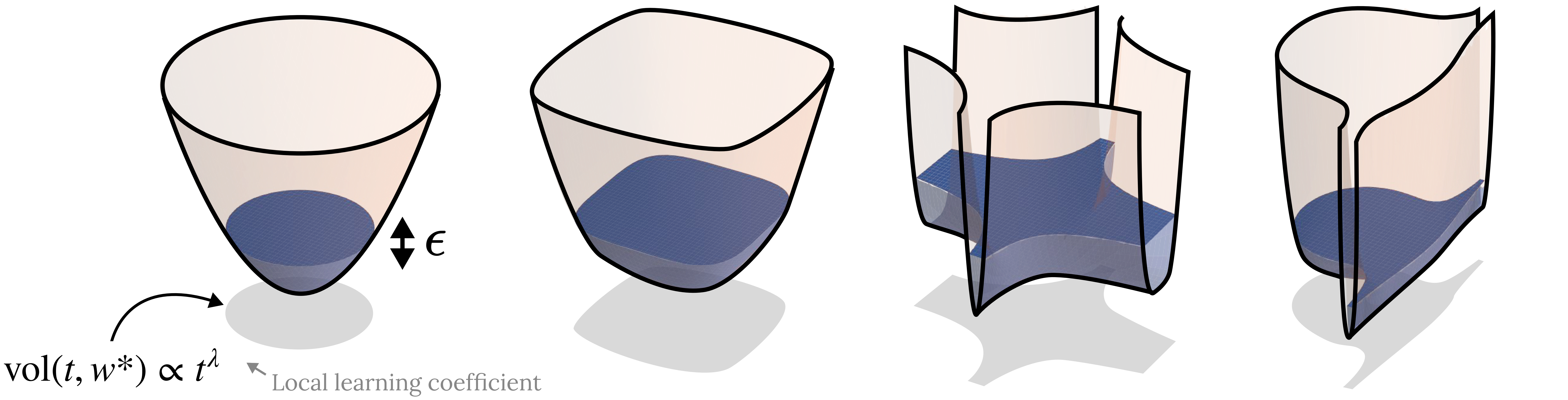
$$\lambda(w^*) := -\lim_{t \to 0} \log_2 \left[\mathrm{vol}(\tfrac{1}{2}t, w^*) / \mathrm{vol}(t, w^*) \right]$$

where for some ball B around w^*

$$\mathrm{vol}(t, w^*) = \int_{|L(w) - L(w^*)| < t,\, w \in B} \varphi(w) dw$$

See Watanabe's books and "The local learning coefficient: a singularity-aware complexity measure" E. Lau, Z. Furman, G. Wang, D. Murfet, S. Wei AISTATS 2025

Examples



$\mathrm{vol}(t, w^*) \propto t^{\lambda}$ Local learning coefficient

$$L(x,y) = x^2 + y^2$$
$$\lambda(0,0) = 1$$
$$\lambda \leq \frac{d}{2}$$

$$L(x,y) = x^4 + y^4$$
$$\lambda(0,0) = \frac{1}{2}$$

$$L(x,y) = x^2 y^4$$
$$\lambda(x,0) = \frac{1}{4}$$

$$L(x,y) = (x-1)^2(x^2+y^2)^4$$
$$\lambda(0,0) = \frac{1}{8}, \ \lambda(1,y) = \frac{1}{2}$$

The LLC can be defined in terms of the rate at which the parameter space volume (within a given neighborhood and with a given maximum loss) shrinks as the loss threshold is reduced to that of the local minimum. We show four population loss landscapes for a two-dimensional parameter space with decreasing LLC (increasing degeneracy). In these examples, the local multiplicity is 1.

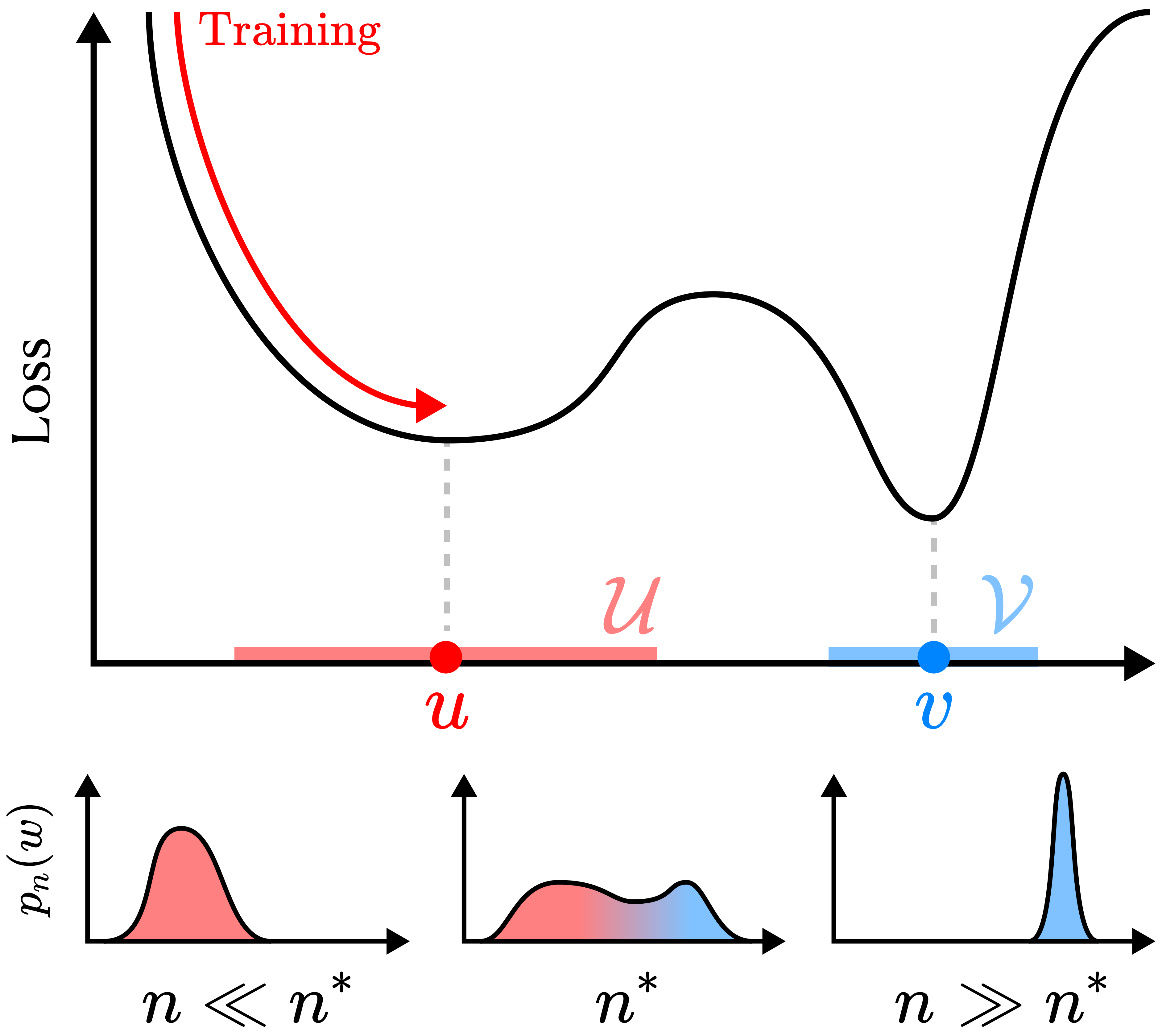
Free Energy Formula

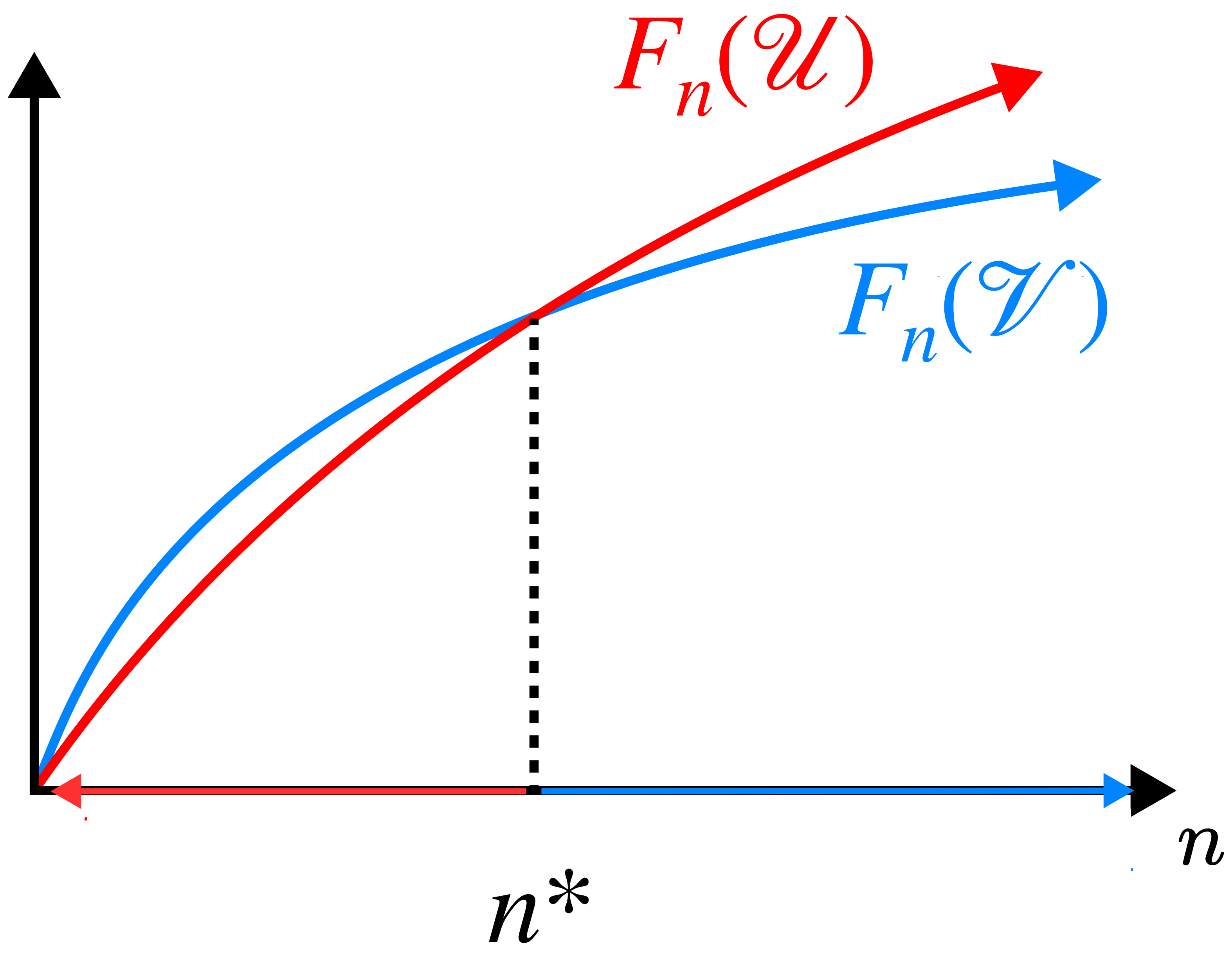
- Assume *relative finite variance* [**Green**, §3.1] in addition to the fundamental conditions of [**Gray**] (excepting realisability) and that there is a point w_0 minimising L in the interior of W.

- **Theorem** (Watanabe): We have by [**Green**, §6.3] for a local semi-analytic neighbourhood $\mathscr{W}$ of a local minimum w^* of L

$$F_n(\mathscr{W}) = nL_n(w^*) + \lambda(w^*)\log n - (m(w^*) - 1)\log\log n + F_n^R + o_p(1)$$

- Here $m \in \mathbb{N}$ is the *multiplicity* and F_n^R is a random variable which converges to a random variable in law.





$$\log \frac{p(v \in \mathcal{V} \mid D_n)}{p(u \in \mathcal{U} \mid D_n)} = F_n(\mathcal{U}) - F_n(\mathcal{V}) \approx \Big(nL_n(u) + \lambda(u)\log n\Big) - \Big(nL_n(v) + \lambda(v)\log n\Big)$$

$$= n \underbrace{\Big(L_n(u) - L_n(v)\Big)}_{>0} + \underbrace{\Big(\lambda(u) - \lambda(v)\Big)}_{<0} \log n$$

Estimating the Learning Coefficient

- The LLC depends on the model (i.e. architecture) and data distribution
- Computing the LLC directly using volumes is intractable
- Theoretically deriving LLCs is difficult (needs nontrivial algebraic geometry in general) and has been done only in a few cases, including DLNs
- However, using theorems of Watanabe we can derive an estimator

$$\hat{\lambda}(w^*) := n\beta\left[\mathbb{E}_{w|w^*,\beta,\gamma} L_n(w) - L_n(w^*)\right]$$

$$p(w \,|\, w^*, \beta, \gamma) \propto \exp\left(-n\beta L_n(w) - \frac{\gamma}{2}\|w - w^*\|_2^2\right)$$

Estimating the Learning Coefficient

$$\hat{\lambda}(w^*) := n\beta\Big[\mathbb{E}_{w|w^*,\beta,\gamma}L_n(w) - L_n(w^*)\Big]$$

$$p(w \,|\, w^*, \beta, \gamma) \propto \exp\Big(- n\beta L_n(w) - \frac{\gamma}{2}\|w - w^*\|_2^2\Big)$$

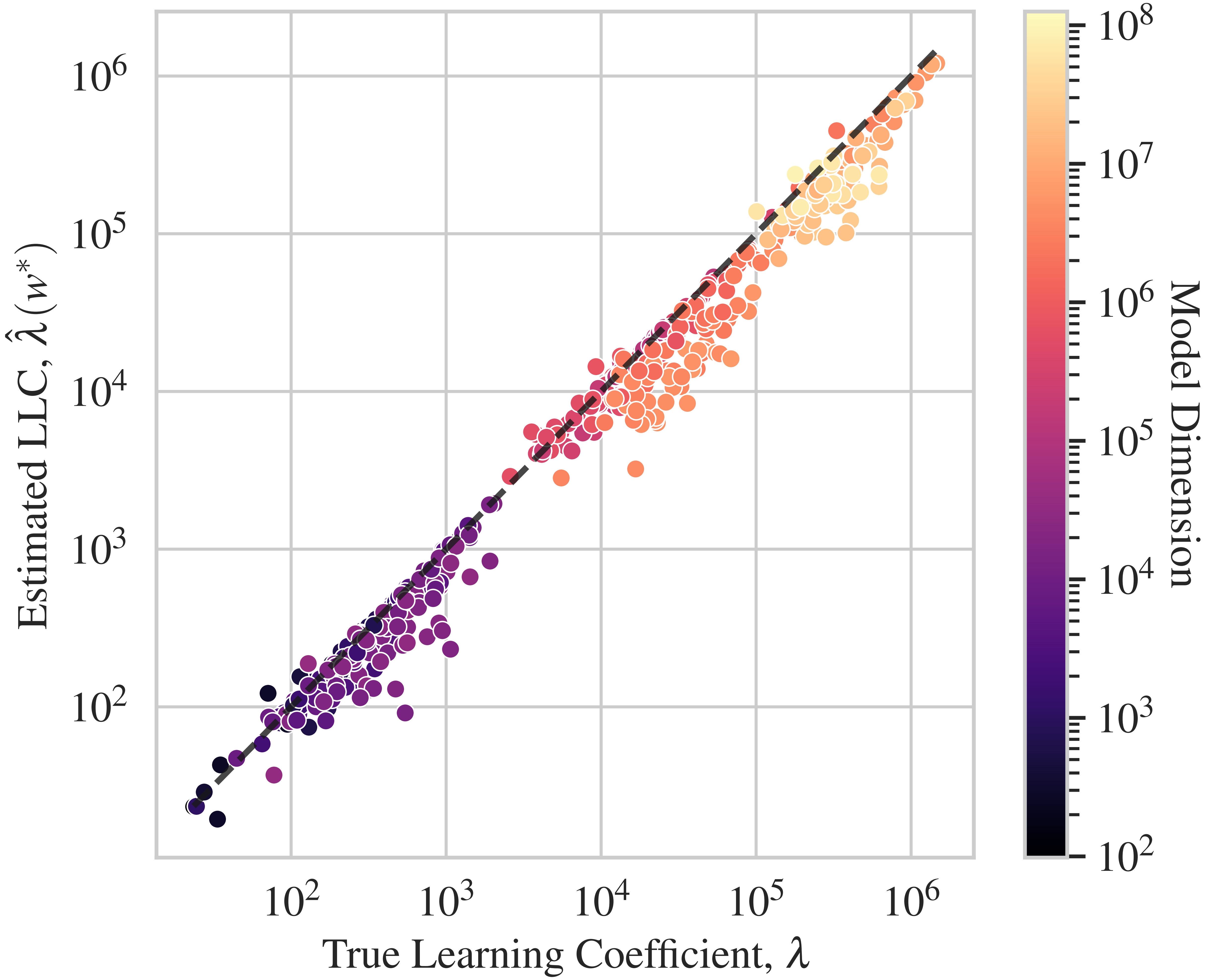
- Sampling from this distribution is difficult
- We use approximate samples based on scalable gradient-based methods (principally SGLD, and an RMSProp variant for larger models)
- Selection of hyper parameters $n\beta, \gamma$ and step-size ϵ and batch-size m for SGLD
- We are *Very Interested* in theoretical and empirical progress on SGLD or related methods as these are the foundation of everything we do

Caveats

- LLC estimation is currently a bit of an art
- We do not trust the absolute value of the estimates (they are far too low in general). However *relative* changes seem meaningful (e.g. over training, or with respect to variations in the data distribution).
- Hyperparameter selection is a bit fraught, but we are pushing on theoretical and empirical fronts to try to understand this better
- We do not believe in accurate posterior approximation with reasonable resources for large neural networks. The hope is that we are able to sample well enough for the expectation values of observables we care about

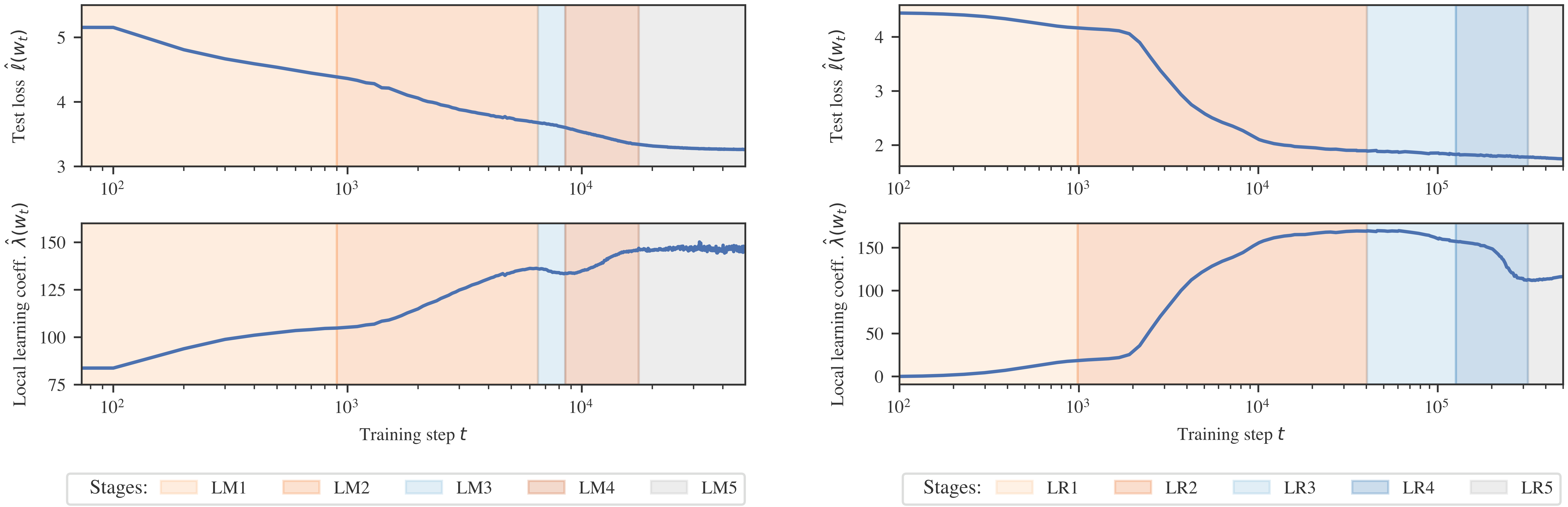
3. Experiments

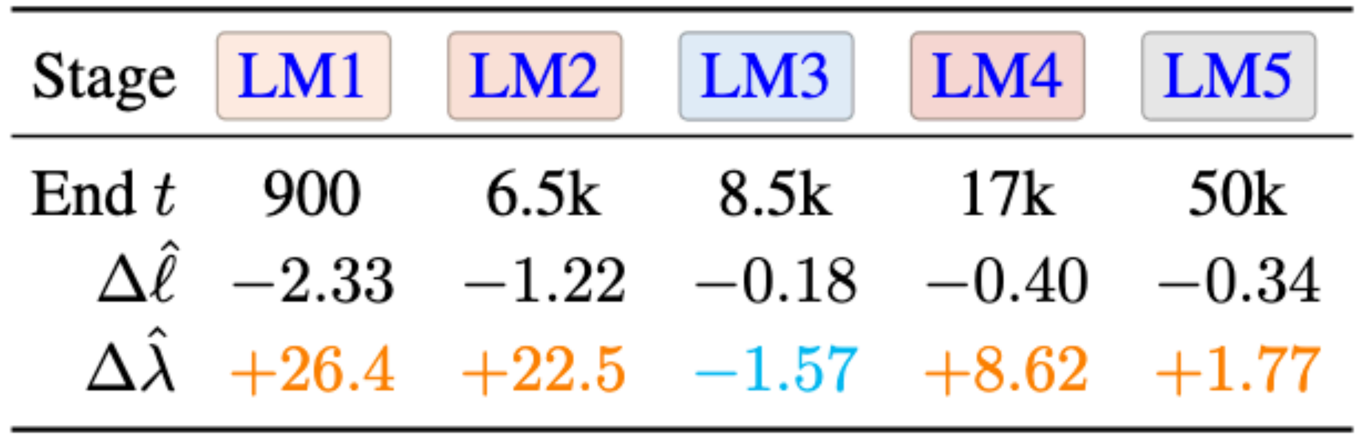
Deep Linear Networks (DLNs)



"The local learning coefficient: a singularity-aware complexity measure"
E. Lau, Z. Furman, G. Wang, D. Murfet, S. Wei AISTATS 2025

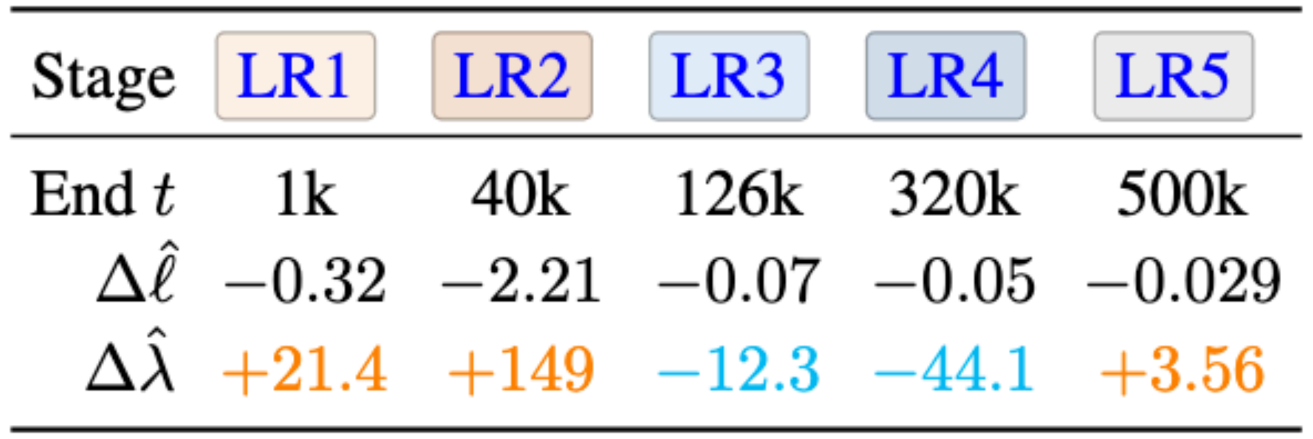
Small Transformers





Stage	LM1	LM2	LM3	LM4	LM5
End t	900	6.5k	8.5k	17k	50k
$\Delta\hat{\ell}$	-2.33	-1.22	-0.18	-0.40	-0.34
$\Delta\hat{\lambda}$	$+26.4$	$+22.5$	-1.57	$+8.62$	$+1.77$

(a) Two-layer attention-only language transformer (LM).



Stage	LR1	LR2	LR3	LR4	LR5
End t	1k	40k	126k	320k	500k
$\Delta\hat{\ell}$	-0.32	-2.21	-0.07	-0.05	-0.029
$\Delta\hat{\lambda}$	$+21.4$	$+149$	-12.3	-44.1	$+3.56$

(b) In-context linear regression transformer (LR).

"Loss landscape geometry drives stagewise development in transformers" J. Hoogland, G. Wang, M. Farrugia-Roberts, L. Carroll, S. Wei, D. Murfet 2024

LLC is “just a number”

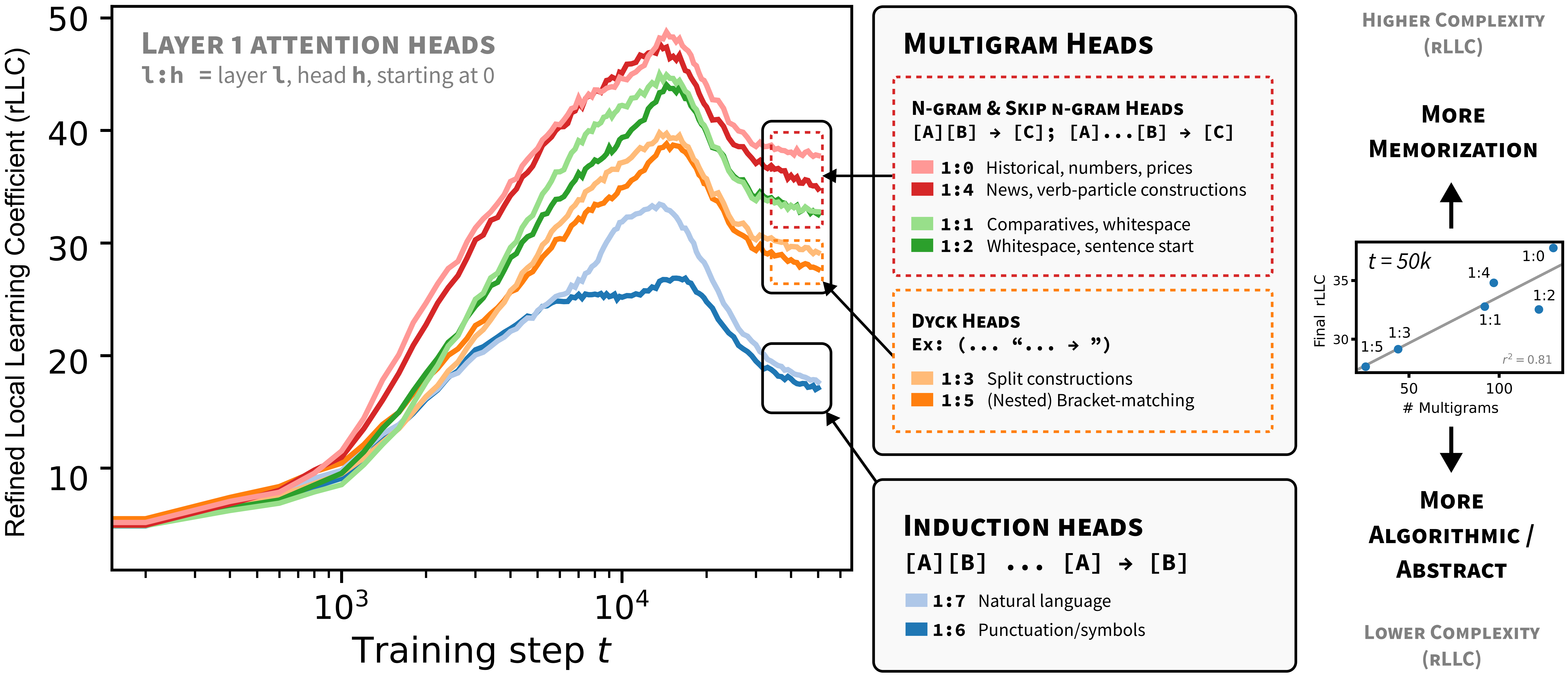
Measure across training

Only allow the parameter to vary in directions within *C*

$$\lambda(w_t\,; C, q')$$

Use *L* for another data distribution

Small Language Models



Multigram Heads

N-gram & Skip n-gram Heads
`[A][B] → [C]; [A]...[B] → [C]`

- **`1:0`** Historical, numbers, prices
- **`1:4`** News, verb-particle constructions
- **`1:1`** Comparatives, whitespace
- **`1:2`** Whitespace, sentence start

Dyck Heads
`Ex: (... “... → ”)`

- **`1:3`** Split constructions
- **`1:5`** (Nested) Bracket-matching

Induction Heads
`[A][B] ... [A] → [B]`

- **`1:7`** Natural language
- **`1:6`** Punctuation/symbols

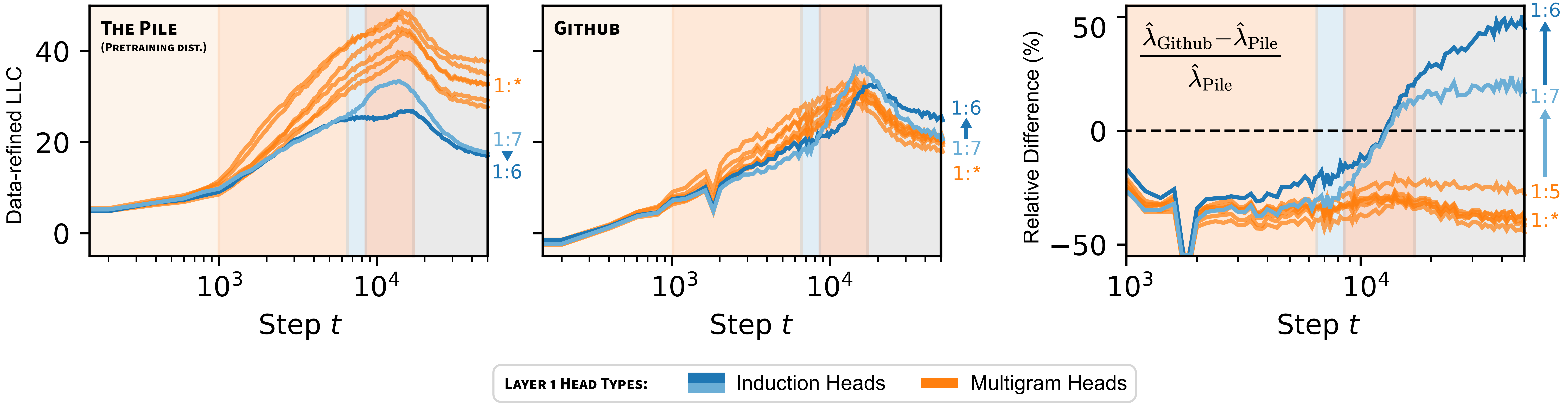
Higher Complexity (rLLC)

More Memorization

More Algorithmic / Abstract

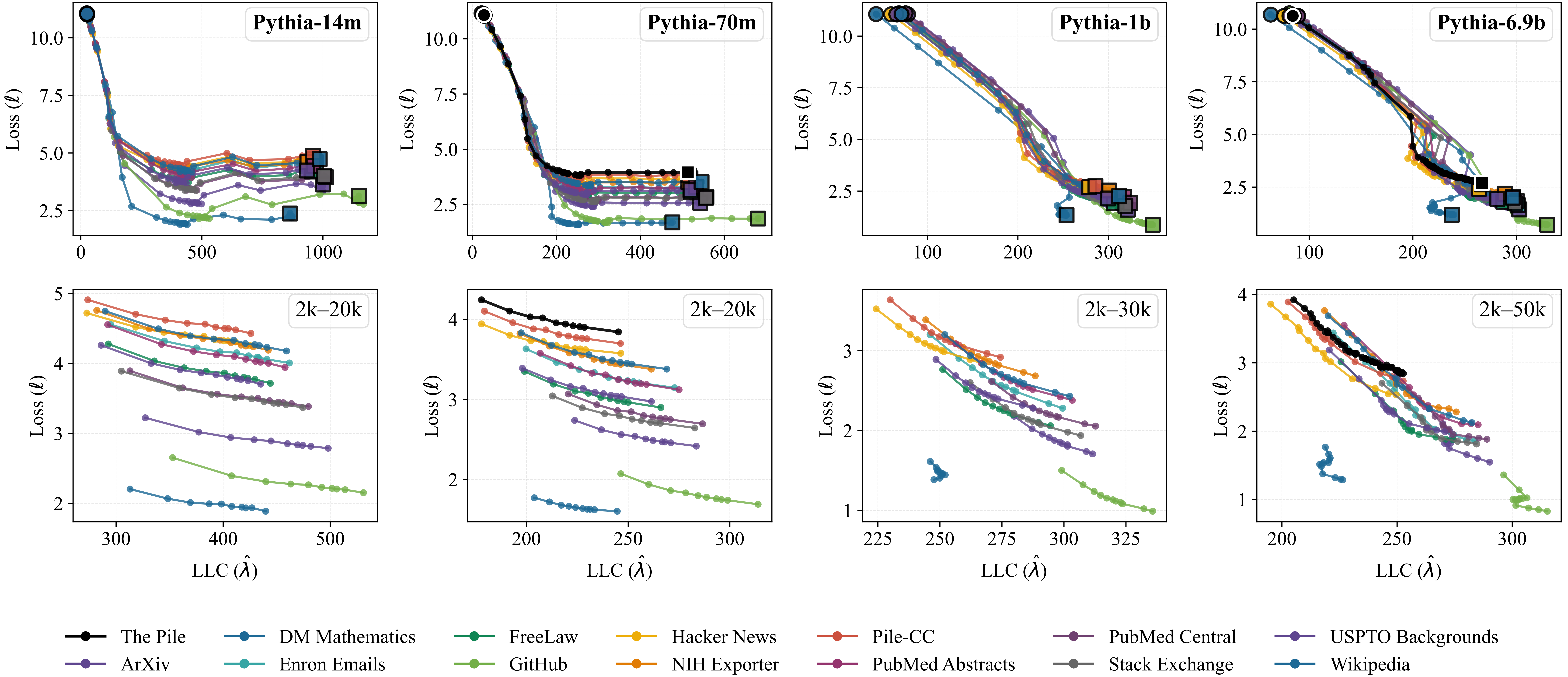
Lower Complexity (rLLC)

“Differentiation and specialization of attention heads via the refined local learning coefficient” G. Wang, J. Hoogland, S. Van Wingerden, Z. Furman, D. Murfet, ICLR 2025 (Spotlight).



Intuition: the LLC has contributions from many “patterns” in the data, with different sensitivity to different patterns, and if you vary the distribution of those patterns different heads respond differently

“Differentiation and specialization of attention heads via the refined local learning coefficient” G. Wang, J. Hoogland, S. Van Wingerden, Z. Furman, D. Murfet, ICLR 2025 (Spotlight).



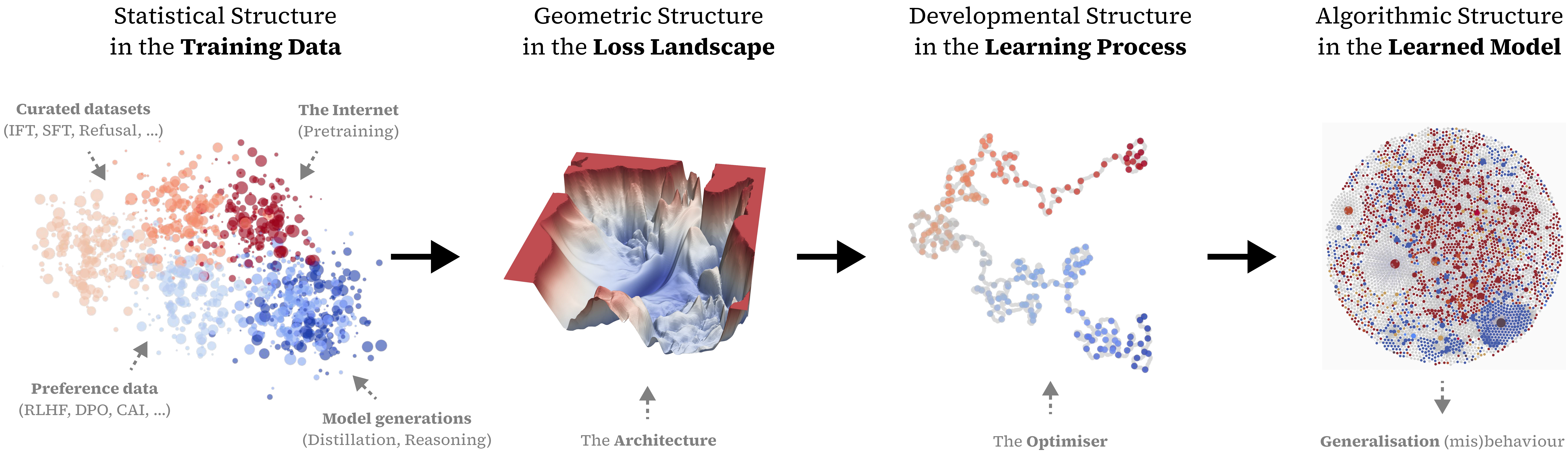
"Psychometrics for Pythia: Connecting evaluations to interpretability using Singular Learning Theory" L. Carroll, A. Reid, J. Hoogland, G. Wang, S. van Wingerden, S. O'Callaghan, D. Murfet, *in preparation*.

4. Structure

Alignment

- The capabilities and alignment of modern AI models are **specified indirectly**, by engineering the data distribution used to train them
- However, model behaviour is **underspecified**: many different ways of performing the computation within the model are consistent with low loss
- Two models with the same evaluation performance but different "modes of computation" may **behave differently** in other settings (e.g. reward hack or not)
- Many approaches to alignment benefit from progress on **understanding these differences**, detecting them and tracing them back to causes
- SLT has non-trivial things to say here, and empirical tools to back it up

"You are what you eat: AI alignment requires understanding how data shapes structure and generalisation" S. Pepin Lehalleur, J. Hoogland, M. Farrugia-Roberts, S. Wei, A. Gietelink Oldenziel, G. Wang, L. Carroll, D. Murfet



Statistical Structure
in the **Training Data**

Curated datasets
(IFT, SFT, Refusal, ...)

The Internet
(Pretraining)

Preference data
(RLHF, DPO, CAI, ...)

Model generations
(Distillation, Reasoning)

Geometric Structure
in the **Loss Landscape**

The **Architecture**

Developmental Structure
in the **Learning Process**

The **Optimiser**

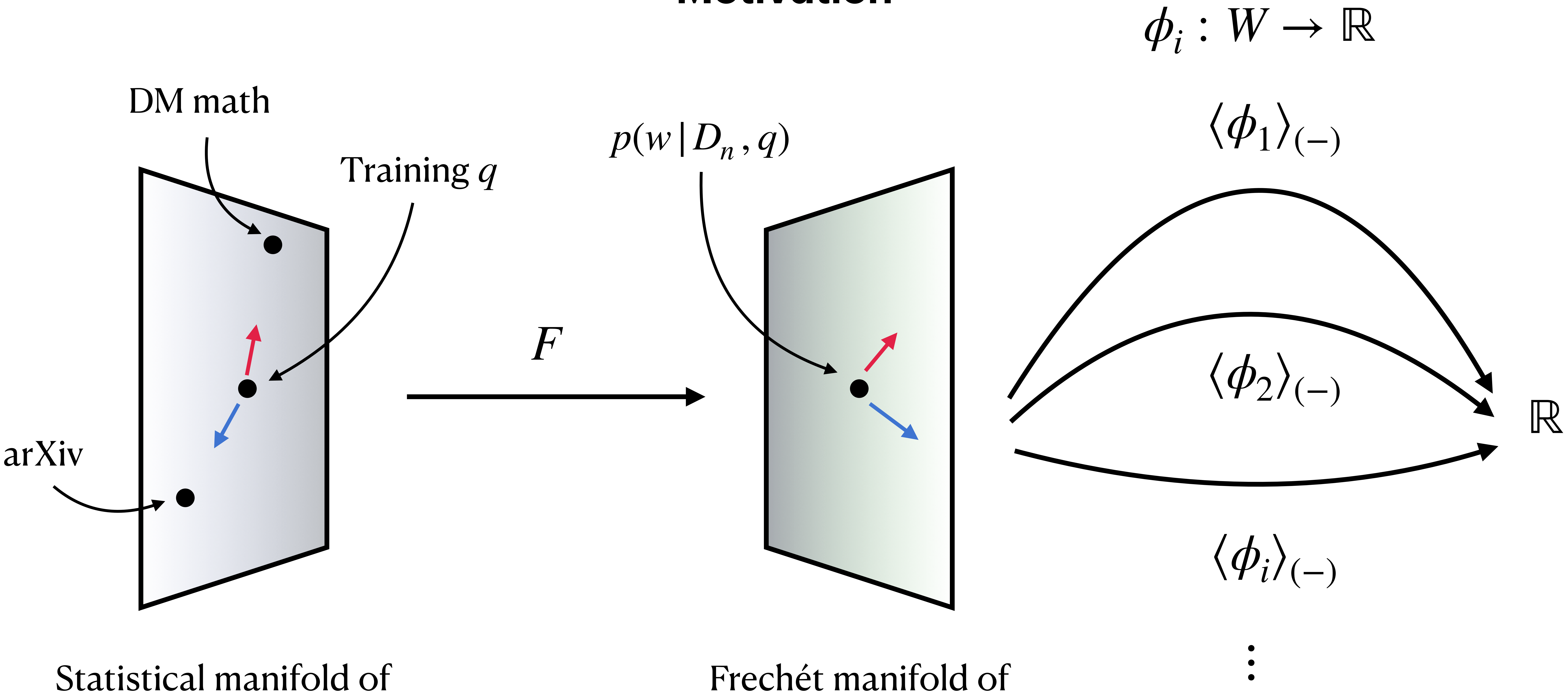
Algorithmic Structure
in the **Learned Model**

Generalisation (mis)behaviour

"You are what you eat: AI alignment requires understanding how data shapes structure and generalisation" S. Pepin Lehalleur, J. Hoogland, M. Farrugia-Roberts, S. Wei, A. Gietelink Oldenziel, G. Wang, L. Carroll, D. Murfet

Susceptibilities

Motivation



$\phi_i : W \to \mathbb{R}$

DM math

Training q

arXiv

F

$p(w \mid D_n, q)$

$\langle \phi_1 \rangle_{(-)}$

$\langle \phi_2 \rangle_{(-)}$

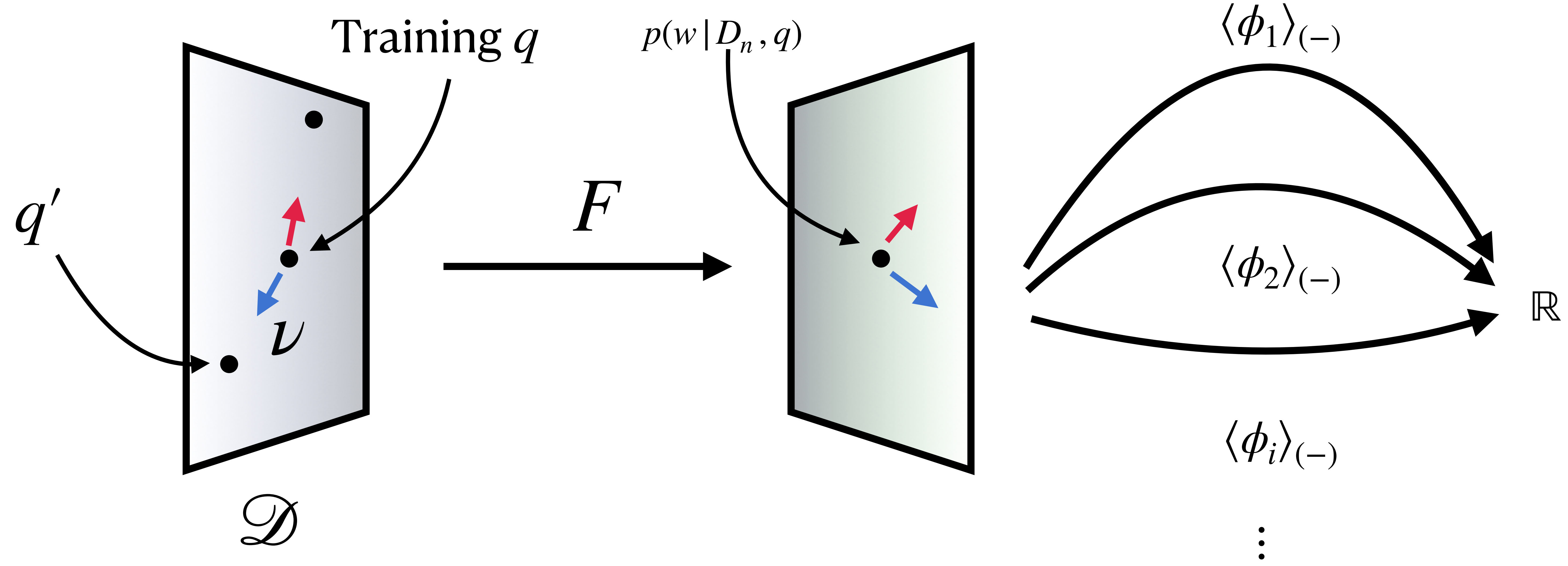
$\langle \phi_i \rangle_{(-)}$

$\mathbb{R}$

$\vdots$

Statistical manifold of
data distributions $\mathscr{D}$

Frechét manifold of
“Posteriors” $\mathrm{Dens}^\infty(W)$

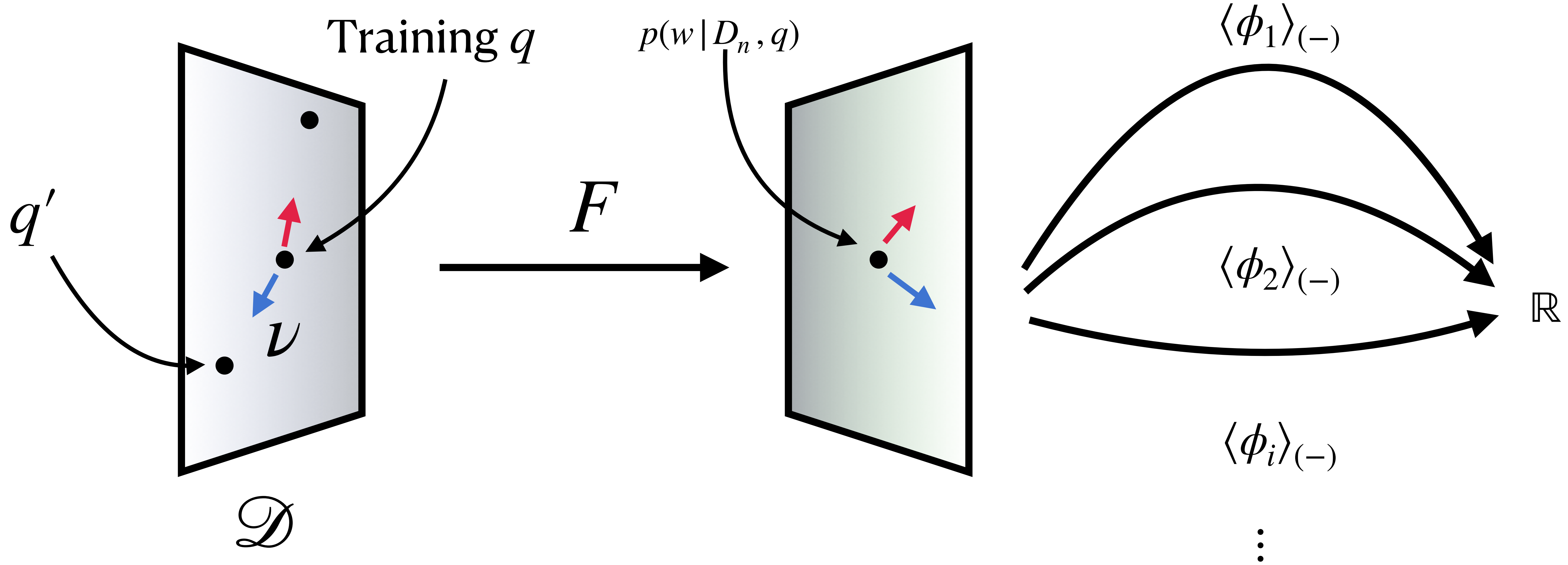

Training q
q'
ν
$\mathscr{D}$
F
$p(w\,|\,D_n, q)$
$\langle \phi_1 \rangle_{(-)}$
$\langle \phi_2 \rangle_{(-)}$
$\langle \phi_i \rangle_{(-)}$
$\vdots$
$\mathbb{R}$

Given a tangent vector $\nu \in T_q\mathscr{D}$ and observable $\phi : W \to \mathbb{R}$ the associated *susceptibility* is

$$\chi(\nu) := \frac{1}{n} T_q\Big(\langle \phi \rangle_{(-)} \circ F\Big)(\nu) \in \mathbb{R}$$

For example, if ν is the infinitesimal form of $h \mapsto q_h := (1-h)q + hq'$

$$\chi = \frac{1}{n}\frac{\partial}{\partial h}\int_W \phi(w)p(w\,|\,D_n^h, q_h)dw\Big|_{h=0}$$


Training q
$p(w \mid D_n, q)$
$\langle \phi_1 \rangle_{(-)}$
q'
F
$\langle \phi_2 \rangle_{(-)}$
$\mathbb{R}$
ν
$\langle \phi_i \rangle_{(-)}$
$\mathscr{D}$
$\vdots$

As with the LLC, we do this *locally* at $w^* \in W$ and for observables that depend on a component C where $W = U \times C$, $w^* = (u^*, c^*)$ and define estimators using SGLD

$$\chi(w^*; C, \nu) = -\operatorname{Cov}\left[\phi_C, \frac{\partial}{\partial h} L_n^h\right] = \frac{1}{n^2\beta} \frac{\partial}{\partial h} \hat{\lambda}(w^*; C, q_h)\Big|_{h=0}$$

$$\phi_C = \delta(u - u^*)\Big[L_n^h(w) - L_n^h(w^*)\Big]$$

Susceptibilities

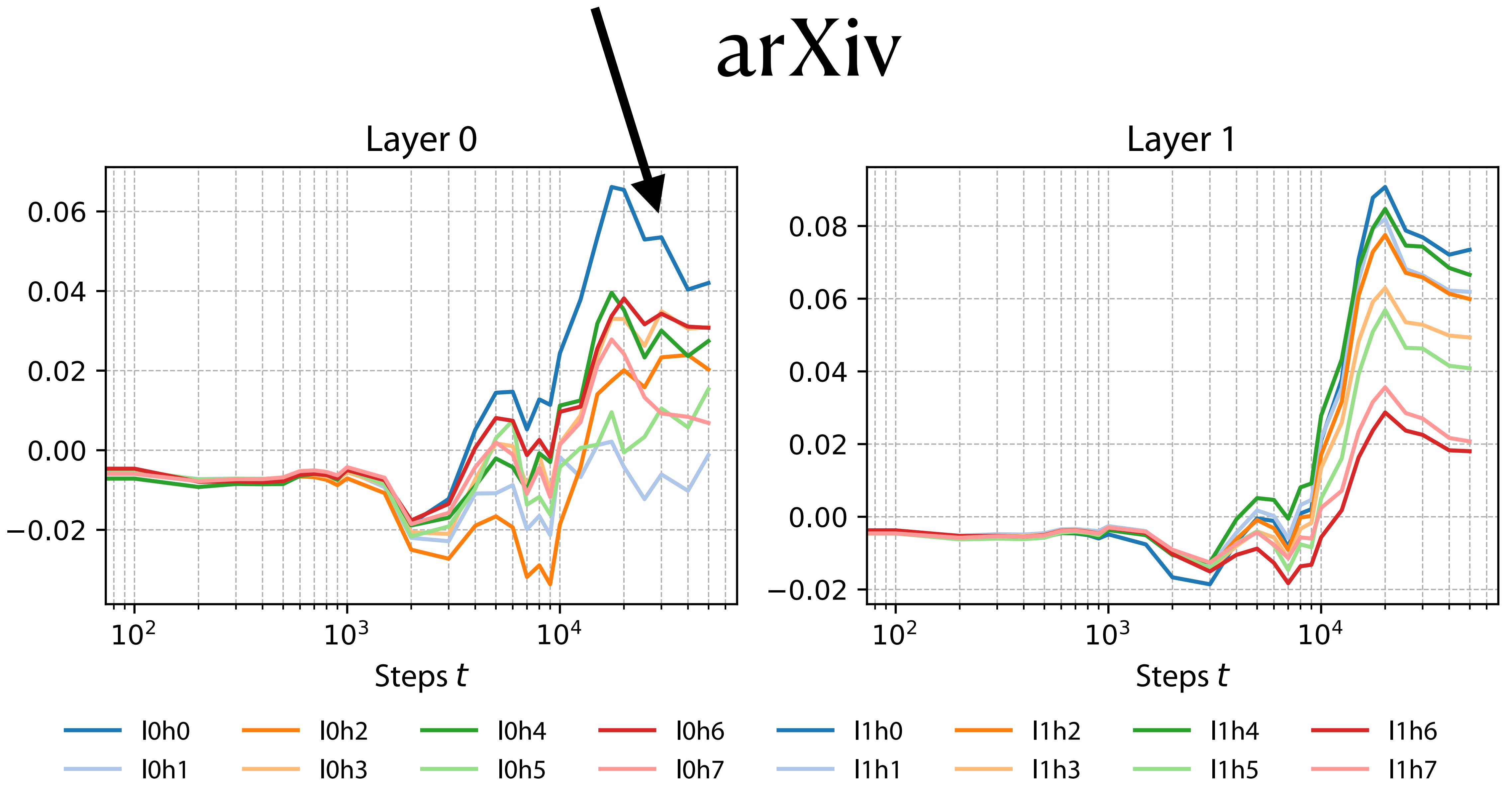
Experiments

- We took a two-layer attention-only transformer trained on the Pile, and studied susceptibilities for pairs consisting of ϕ the above observable associated to individual attention heads, and ν the variation of the Pile in the direction of one of its subsets (e.g. arXiv, NIH exporter).
- These numbers (which can be positive or negative) tracked across training checkpoints give a picture of the degree to which varying the data distribution (e.g. to include more math or code) pushes the posterior restricted to parts of the model
- This is work in progress ᘔ

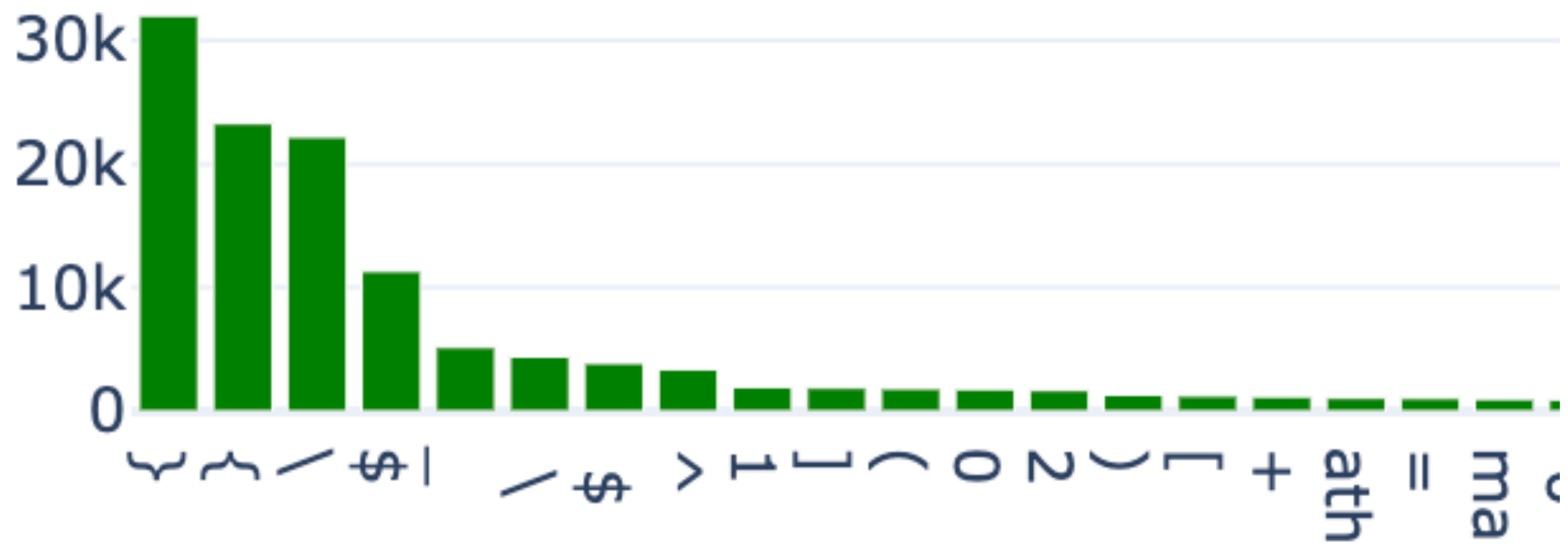
“Studying small language models with susceptibilities” G. Baker, G. Wang, J. Hoogland, D. Murfet, *in preparation*

Layer 0 head 0

arXiv

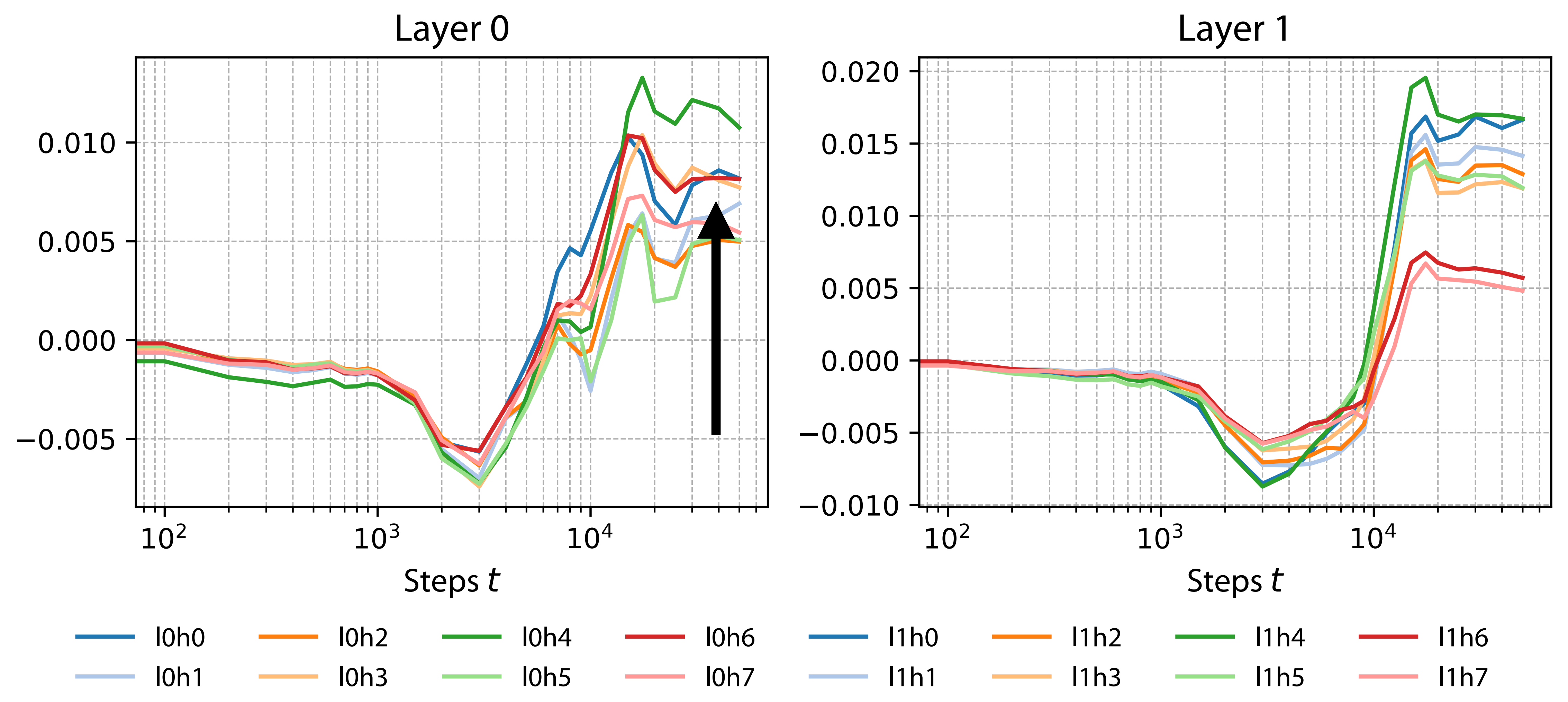


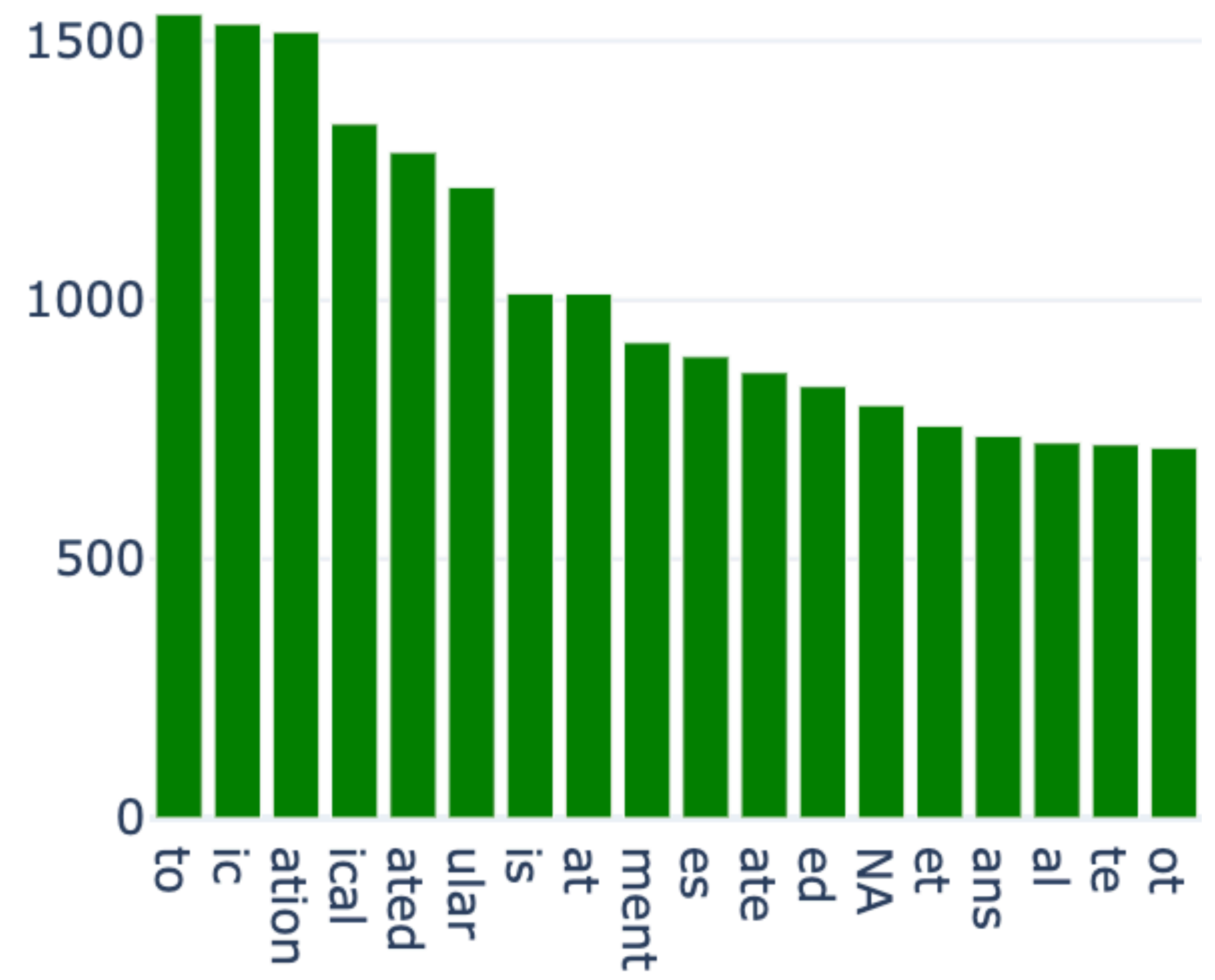
Top contributing +ve tokens



Mean susceptibility of occurrences of } is 4.5

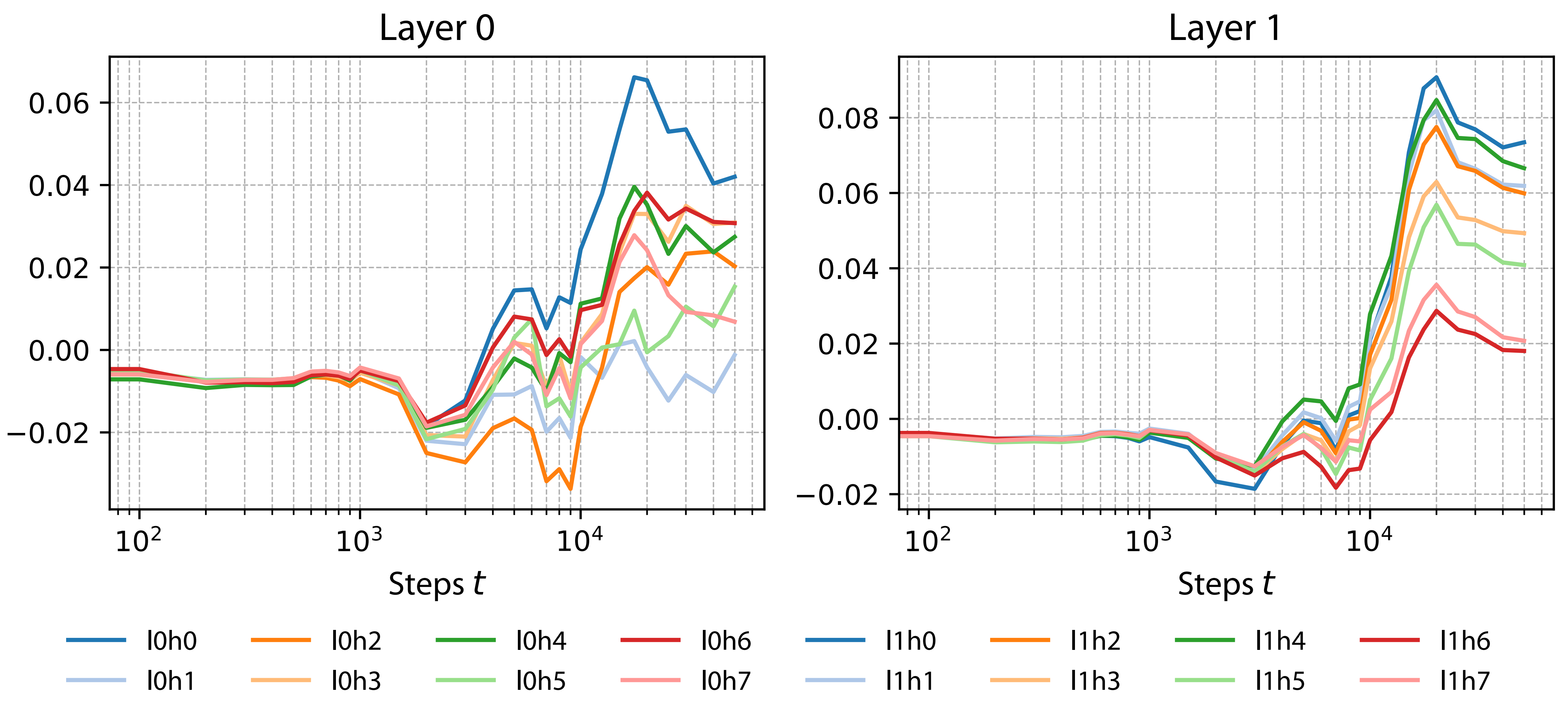
NIH exporter



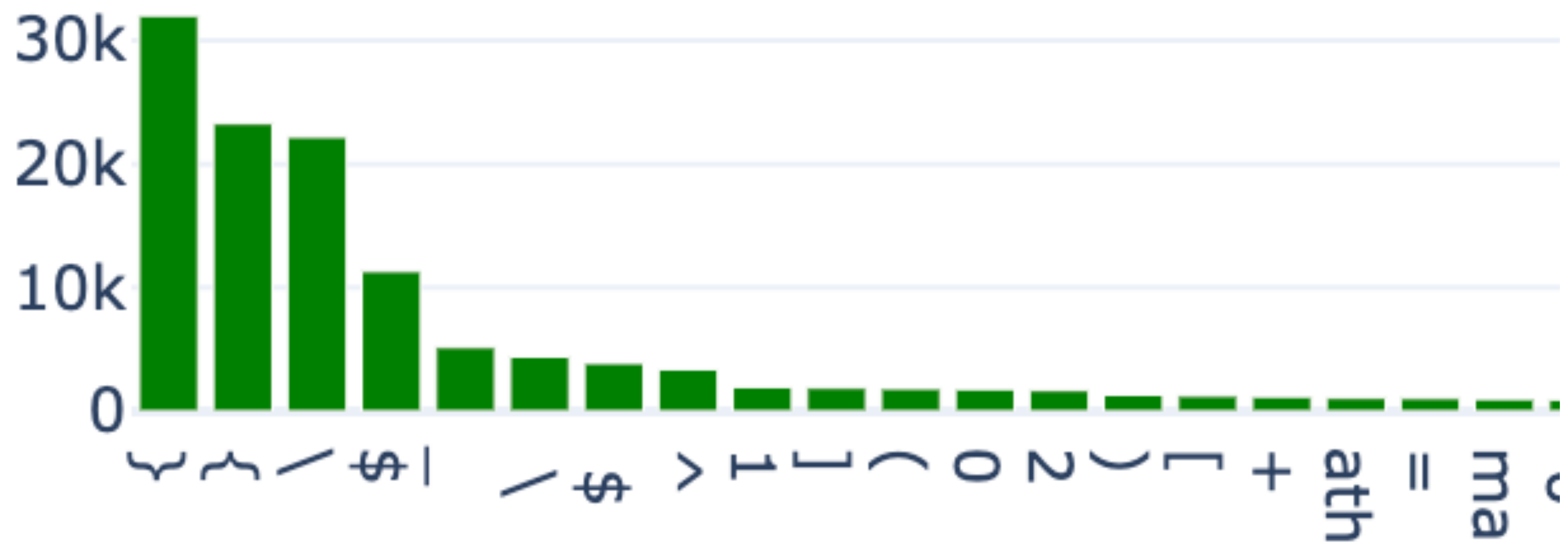


Mean susceptibility of occurrences of to is 1.0

arXiv

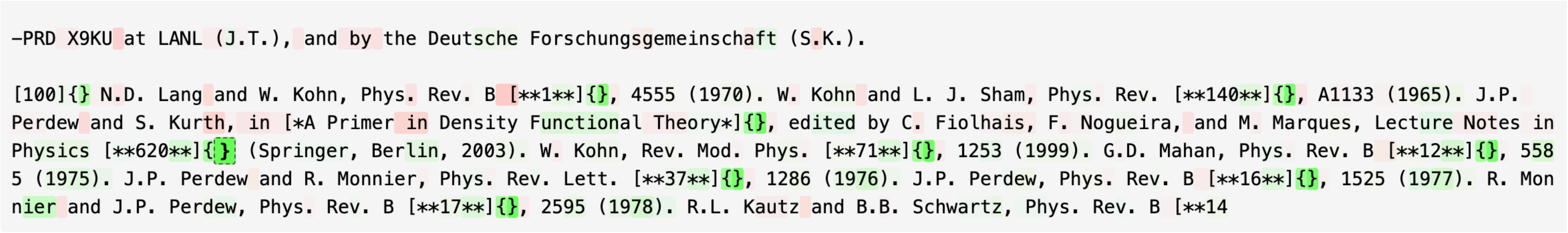


Top contributing +ve tokens



Mean susceptibility of occurrences of } is 4.5

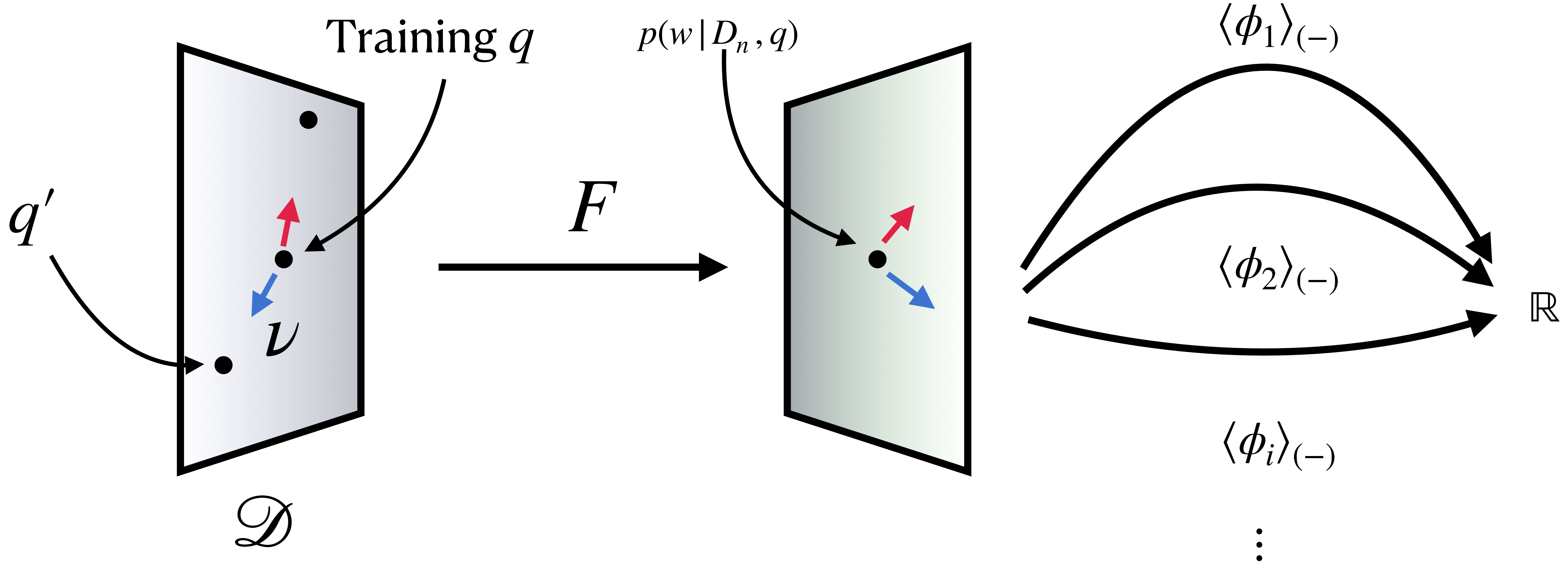
Per-token susceptibilities on an arXiv sample (green is positive)



```
-PRD X9KU at LANL (J.T.), and by the Deutsche Forschungsgemeinschaft (S.K.).

[100]{} N.D. Lang and W. Kohn, Phys. Rev. B [**1**]{}, 4555 (1970). W. Kohn and L. J. Sham, Phys. Rev. [**140**]{}, A1133 (1965). J.P.
Perdew and S. Kurth, in [*A Primer in Density Functional Theory*]{}, edited by C. Fiolhais, F. Nogueira, and M. Marques, Lecture Notes in
Physics [**620**]{} (Springer, Berlin, 2003). W. Kohn, Rev. Mod. Phys. [**71**]{}, 1253 (1999). G.D. Mahan, Phys. Rev. B [**12**]{}, 558
5 (1975). J.P. Perdew and R. Monnier, Phys. Rev. Lett. [**37**]{}, 1286 (1976). J.P. Perdew, Phys. Rev. B [**16**]{}, 1525 (1977). R. Mon
nier and J.P. Perdew, Phys. Rev. B [**17**]{}, 2595 (1978). R.L. Kautz and B.B. Schwartz, Phys. Rev. B [**14
```

Conjecture: Much of the information "in" Layer 0 head 0 is about brackets


Training q
$p(w \mid D_n, q)$
$\langle \phi_1 \rangle_{(-)}$
q'
F
$\langle \phi_2 \rangle_{(-)}$
$\mathbb{R}$
ν
$\langle \phi_i \rangle_{(-)}$
$\mathscr{D}$
$\vdots$

- **Interpretability:** by studying the data matrix $\left\{\hat{\chi}(w^*; C_i, \nu_j)\right\}_{i,j}$ we hope to understand internal structure in large models and the effect of training and fine-tuning on that structure
- **Patterning:** if we can detect sensitivity of the posterior to certain patterns in the data, we may be able to more effectively steer the learning process (both in pre-training and in post-training)

Conclusion

- SLT is a mathematical theory of Bayesian statistics, due to Sumio Watanabe
- We have developed and tested a scalable estimator of the core quantity, the LLC, to transformer models of 7B parameters.
- Across a variety of settings it reflects changes in complexity (for parts of the model and parts of the data) that we can independently validate “make sense”
- The LLC and related quantities like susceptibilities provide a foundation for understanding how **structure in data shapes structure in models**
- These theoretical and empirical foundations can be applied to many problems in AI safety, including but not limited to interpretability

timaeus.co