UNIVERSITY OF OSLO

Schibsted

BoRA: Bayesian Hierarchical Low-Rank Adaptation for Multi-Task Large Language Models

Simen Eide, University of Oslo, Schibsted Media
Arnoldo Frigessi, University of Oslo

Arxiv

Motivation

For multiple related LLM finetuning tasks, should you:

- Train each tasks separately with **independent parameters**?
- Train all tasks together with **shared parameters**?

We present a third alternative:

Train all LLM finetuning tasks with separate (LoRA)-parameters using hierarchical priors to share information between tasks

Method

Our approach extends the Low-Rank Adaptation (LoRA) technique by introducing a Bayesian hierarchical model. For each task d, we define task-specific low-rank parameters θ_d, and use a global hierarchical prior Θ to capture shared structures.

The likelihood for task d is expressed as:

$$L(\mathcal{D}_d|\theta_d) = \prod_{n=1}^{N_d} \prod_{i=1}^{W_n-1} \text{LLM}(w_{i+1}|w_{1:i};\theta_d)$$

where $\mathcal{D}_d$ represents the dataset for task d.

The hierarchical prior is modeled as:

$$P(\theta_{1:D}|\Theta,\tau) = \prod_{d=1}^{D} \mathcal{N}(\theta_d;\Theta,\frac{1}{\tau}I)$$

where τ controls the strength of the prior.

The posterior distribution combines the likelihood and the prior:

$$P(\theta_{1:D}|\mathcal{D},\tau) \propto \prod_{d=1}^{D} L(\mathcal{D}_d|\theta_d)P(\theta_d|\Theta,\tau)$$

Optimization is performed jointly over the task-specific parameters $\theta_{1:D}$ and the global parameters Θ.

Model Interpretation

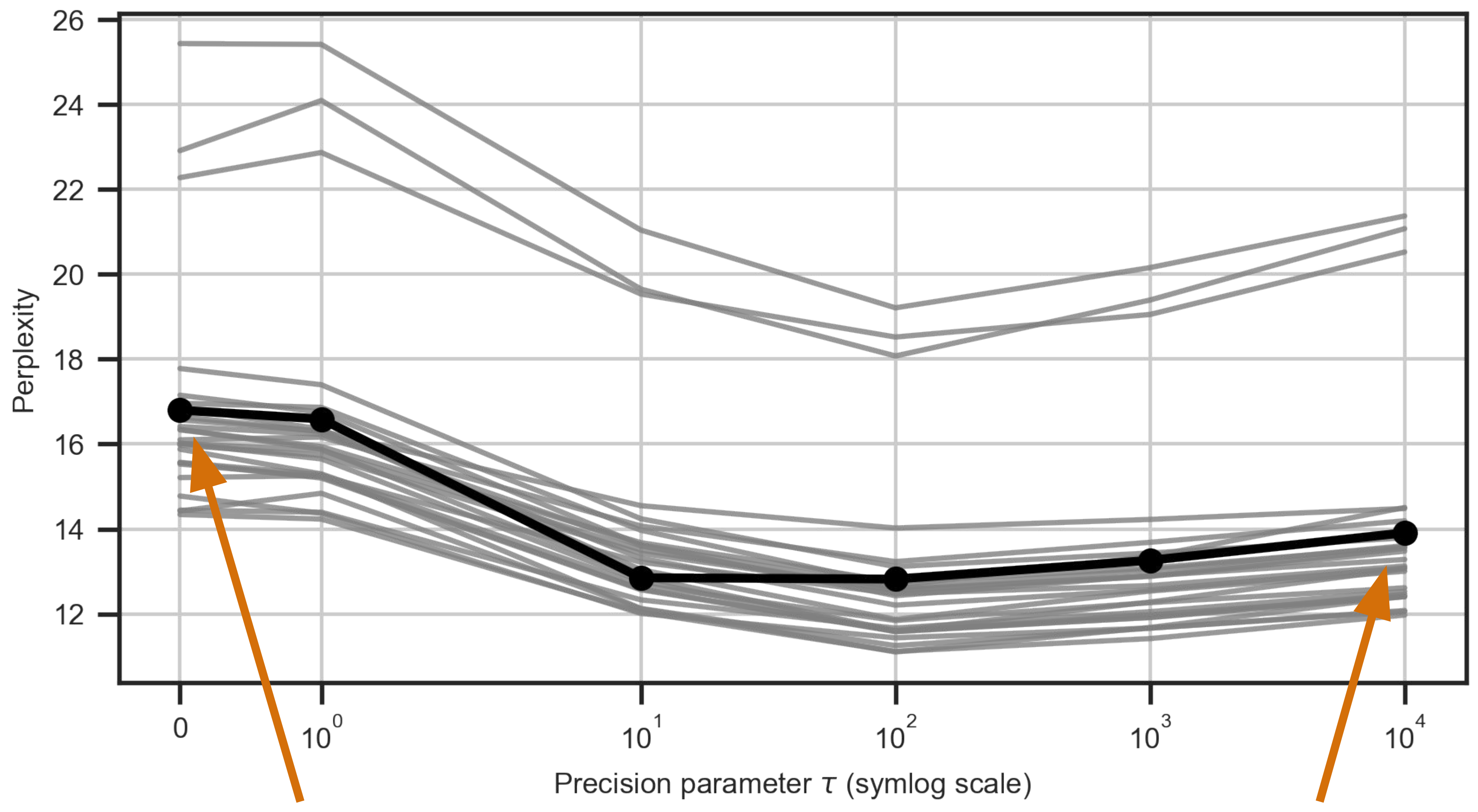
- Each task has a separate set of LoRA parameters
- A global set of LoRA parameters functions as a hierarchical prior: effectively a “weighted average” of all individual task parameters
- τ controls the closeness of task parameters to the global parameters
- Generalisation of the two baseline methods:
  τ=0: Shared parameters
  τ→∞: Independent parameters

Experiment

- We show the method on the Talk of Norway dataset: a collection of speeches from Norwegian parliament speakers from different parties and dialects
- Task: next token generation
- Each speaker is a task, each speech is a document

Results

- BoRA achieved **lower perplexity (higher likelihood)** on test set across tasks compared to both baselines
- Tasks with smaller datasets borrows (successfully) more information from hierarchical prior



Shared parameters (τ=0)

Independent parameters (τ → ∞)

Figure: The thick black line is the overall test perplexity across all tasks, and the thinner grey lines represent the test perplexity for each individual task. The leftmost point corresponds to training each task independently (τ = 0), and the rightmost point corresponds to the limiting case when all task-specific model parameters are constrained to be equal (τ → ∞).

Benefits (why BoRA?)

- Automatically borrows information from related tasks
- Tasks with smaller datasets borrow more from hierarchical prior
- Improved finetuning

Future Work

- How to determine τ (how much information to share / how similar the tasks are)
- Extend the method to other finetuning approaches (DoRa, MoRA, LoRA+, …)
- Evaluate on more datasets