MESS+: Energy-Optimal Inferencing in Language Model Zoos with Service Level Guarantees

Ryan Zhang[1], Herbert Woisetschläger[2], Shiqiang Wang[3], Hans-Arno Jacobsen[4]
1 - Horace Greeley High School, 2 - Technical University of Munich, 3 - IBM Research, 4 - University of Toronto

Research Question: *Can we select appropriate models from the model zoo to ensure energy efficiency while satisfying service level agreements (SLAs)?*

1 Introduction

- *Deep learning infrastructure providers* and *end users* are confronted with an abundance of models (**model zoo**) for language modeling tasks.
- Related to two major research directions: **Dynamic Inference** and **Inference Request Scheduling**.
- Our work, MESS+, automatically selects a readily pre-trained model.
- Since *MESS+* routes inference requests to different models, it can build on top of existing scheduling techniques.

We are solving a Tri-fold Problem

1. End-users primarily care about *correct model output*
2. Inference endpoint providers prioritize *low operating costs*
3. Enterprise use-cases require consistent high quality model output while keeping costs in check through **Service Level Agreements** (SLAs)

2 Methodology

Algorithm 1: Selecting the Model with Energy-optimal Service level GuaranteeS (MESS+)

Input: T; V; α; c; $\{E_m(t) : \forall m, t\}$; learning rate $\eta > 0$
Output: Outputs of models chosen for all t

```
Initialize Q(1) ← 0; predictor parameters x_m to a common
  random vector for all m; k ← 1 ;
for t ← 1 to T do
    Compute p_t ← min(1, c / ∛t);
    Sample X_t ~ Bernoulli(p_t);
    if X_t = 1 then
        // Exploration
        foreach m ∈ {1, 2, ..., M} do
            Obtain true accuracy A_m(t);
            x_{m,t+1} ←
              x_{m,t} − η∇_x (Â(x_{m,t}, a_t) − A_m(t))^2;
        m* ← arg max_m A_m(t);
    else
        m* ← arg min_m V·E_m(t) + Q(t)·(α − Â_m(t));
        x_{m,t+1} ← x_{m,t};
    Get output from model m* and its accuracy A_{m*}(t);
    // Virtual queue update
    Q(t+1) ← max{0, Q(t) + α − A_{m*}(t)};
```

Exploration probability over time with cube root decay

Decision problem for each request

A virtual queue is used to capture SLA violations

Overall Control Problem:

- ***E*** Energy consumption (joules) for a model.
- ***A*** Accuracy of a model's response (requires feedback signal).
- ***y*** Binary variable; 1 if a model is chosen.
- ***α*** SLA minimum accuracy requirement.

$$\min_{\{y_m(t):\forall t,m\}} \quad \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} y_m(t)E_m(t),$$
$$\text{s.t.} \quad \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} y_m(t)A_m(t) \geq \alpha,$$
$$\sum_{m=1}^{M} y_m(t) = 1, \forall t \in \{1,\ldots,T\},$$
$$y_m(t) \in \{0,1\}, \forall t, m,$$

Challenges:

- Objective and constraints are correlated over time
- Characteristics of future requests cannot be predicted

3 Accuracy Predictor

MSE Objective for accuracy predictor:

$$L(\mathbf{x}_m) = \mathbb{E}_{\mathbf{a}_t}\left(\hat{A}(\mathbf{x}_m, \mathbf{a}_t) - A_m(t)\right)^2$$

To predict the accuracy, we train a predictor for *K* steps. The convergence upper bound is:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla L(\mathbf{x}_{m,t_k})\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$$

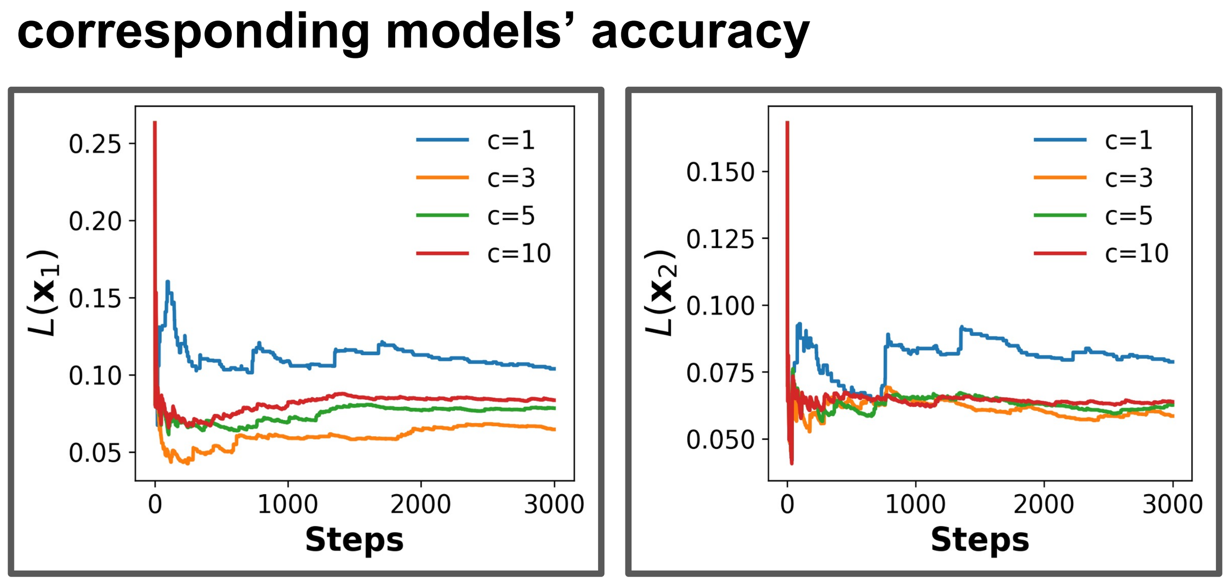
SGD convergence bound for our choice of p_t:

$$\mathbb{E}[K] = \Theta\left(T^{\frac{2}{3}}\right) \qquad \mathcal{O}\left(1/\sqrt[3]{T}\right)$$

Upper bound of *avg. add'l* energy consumption: $\frac{1}{T}\sum_{k=1}^{K} E = \frac{KE}{T}$

For our choice of p_t: $\mathcal{O}\left(E/\sqrt[3]{T}\right)$

Our linear classifiers rapidly learn to predict their corresponding models' accuracy



4 MESS+ ensures energy minimal SLA compliance

- ***MESS+*** satisfies the SLA with requirement *α* while consuming the least energy among all compliant strategies
- Analyzing V reveals that its optimal value is inversely related to the minimum service requirement *α*
- Balancing level of exploration is important as large *c* implies more training of $\mathbf{x}_m$ but can also lead to overfitting on incoming requests

MESS+ reduces the energy consumption of an inference service by up to 2.5x

Model	WMT14 ($\alpha = 0.52$)			CNNDailyMail ($\alpha = 0.315$)		
	Accuracy (BLEU)	Energy (Joules)	Meets α	Accuracy (ROUGE1)	Energy (Joules)	Meets α
TinyLlama	49.1 ± 0.6	44.639 ± 0.6	No	30.9 ± 0.3	142.080 ± 1.4	No
Llama-2 13B	55.1 ± 0.4	527.870 ± 5.1	Yes	32.2 ± 0.3	750.285 ± 7.5	Yes
Random with constraint	52.0 ± 0.0	280.426 ± 3.0	Yes	31.5 ± 0.0	416.466 ± 4.3	Yes
MESS+ ($V = 0.1$, $c = 3$)	52.2 ± 0.2	**149.399** ± 1.3	Yes	31.5 ± 0.1	**163.836** ± 1.5	Yes