Gaussian Process Thompson sampling for Bayesian optimization of dynamic masking-based language model pre-training

Iñigo Urteaga[1,2] Moulay-Zaïdane Draïdia[2] Tomer Lancewicki[3] Shahram Khadivi[4]

inigo.urteaga@columbia.edu mad2314@columbia.edu tomer.lancewicki@walmart.com skhadivi@ebay.com

[1]Applied Physics and Applied Mathematics, Columbia University, NY, USA [2]Data Science Institute, Columbia University, NY, USA

[3]Walmart Global Tech, USA —Work done while at eBay Inc. San Jose, CA. USA [4]eBay Inc., Aachen, Germany

Motivation

- Transformer-based language model (TLM) pre-training is computationally very expensive
- There are many unresolved TLM pre-training choices, e.g., hyperparameter selection
- TLM pre-training hyperparameters are critical, yet search demands time and resources

TLM pre-training as a sequential decision process

- TLM pre-training hyperparameters ψ as a bandit's arms: $a_t = \psi_t$
- TLM pre-training validation losses fit with a Gaussian Process reward model
- Thompson sampling policy to sequentially maximize cumulative rewards

MLM pre-training

- The masked-language model (MLM) loss as objective
- *Random* dynamic masking, with masking choices as hyperparameters
- The MLM objective's dependence with respect to hyperparameters is complex and unknown
- Empirical evaluations of the objective are attainable, i.e., averaged MLM loss in the validation set

Gaussian Process Thompson Sampling (GP-TS) for online optimization of TLM pre-training

Require: TLM and training corpus
Require: Pre-training hyperparameter space Ψ
Require: Number of bandit pre-training interactions T
Require: Number of updates per-interaction u
Require: GP priors $\mu(\cdot)$ and $k(\cdot,\cdot)$ with hyperparameters θ_0

Initialize $\mathscr{A}=\Psi$, $\hat{\theta}_1=\theta_0$, $\mathscr{H}_1=\emptyset$
for $t=1,\cdots,T$ **do**
  Draw posterior sample from the up-to-date GP, i.e., $\mu_a^{(t)} \sim f(\mu_t(a|\hat{\theta}_t), k_t(a,a'|\hat{\theta}_t))$.
  Select next arm based on drawn posterior sample, i.e., $a_t = \mathrm{argmax}_{a'\in\mathscr{A}}\, \mu_{a'}^{(t)}$.
  Run TLM pre-training for u steps, with hyperparameters $\psi_t = a_t$.
  Compute averaged validation loss of pre-trained TLM, i.e.,

$$\bar{y}_t(D_{val};\psi_t) = -\sum_{d\in D_{val}} \frac{\sum_{l_d=1}^{L_d} m_{l_d} \log p(l_d|\widehat{l_d};w,\psi_t)}{\sum_{l_d=1}^{L_d} m_{l_d}} .$$

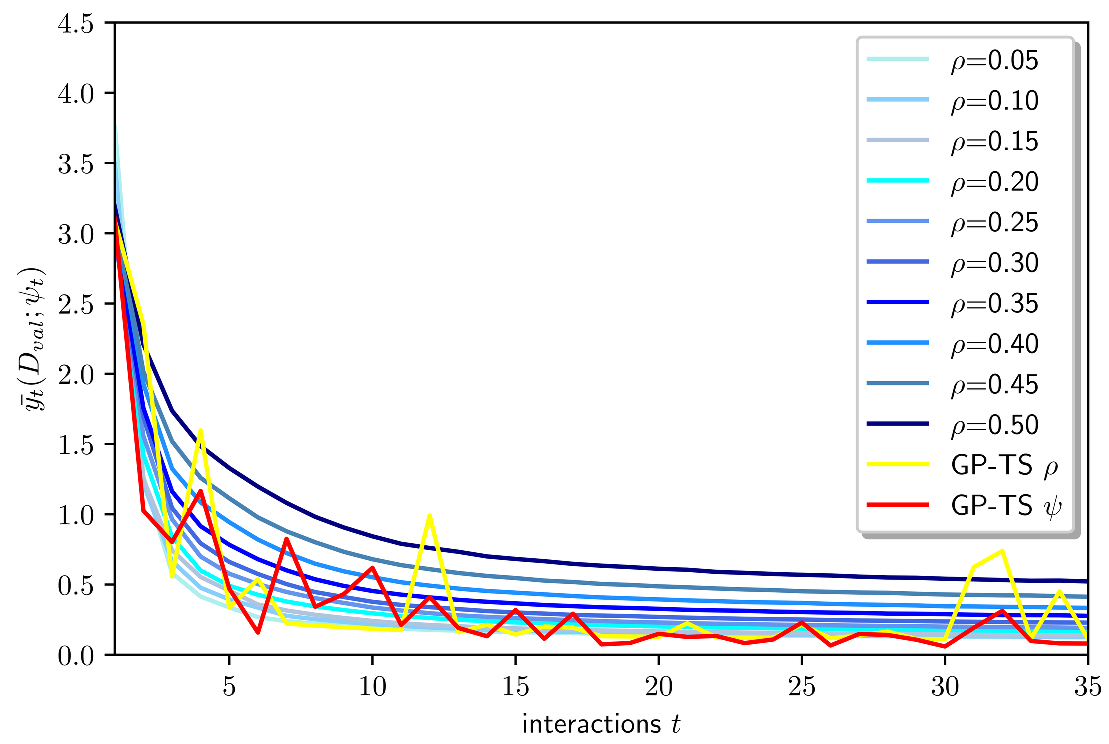
  Observe bandit reward, i.e.,

$$r_t(a_t=\psi_t) = \frac{[-\bar{y}_t(D_{val};\psi_t)]-[-\bar{y}_{t-1}(D_{val};\psi_{t-1})]}{[-\bar{y}_{t-1}(D_{val};\psi_{t-1})]} .$$

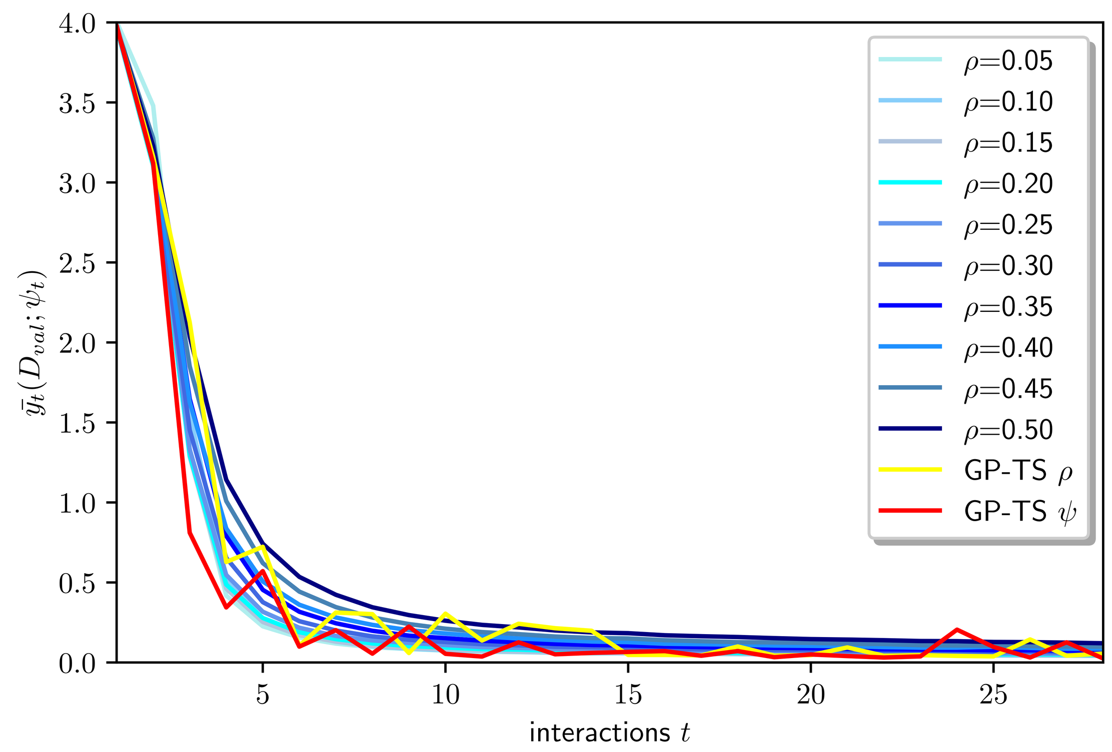
  Update bandit history $\mathscr{H}_{1:t} = \mathscr{H}_{1:t-1} \cup \{a_t, r_t\}$.
  Fit GP model of rewards $r_t(a_t) = f(a_t;\theta)+\epsilon_t$, based on available bandit data $\mathscr{H}_{1:t}$, i.e., $\hat{\theta}_{t+1} = \mathrm{argmax}_\theta \log p(r_{1:t}|f(a_{1:t}),\theta)$.
end for

Pre-training RoBERTa models from *scratch*

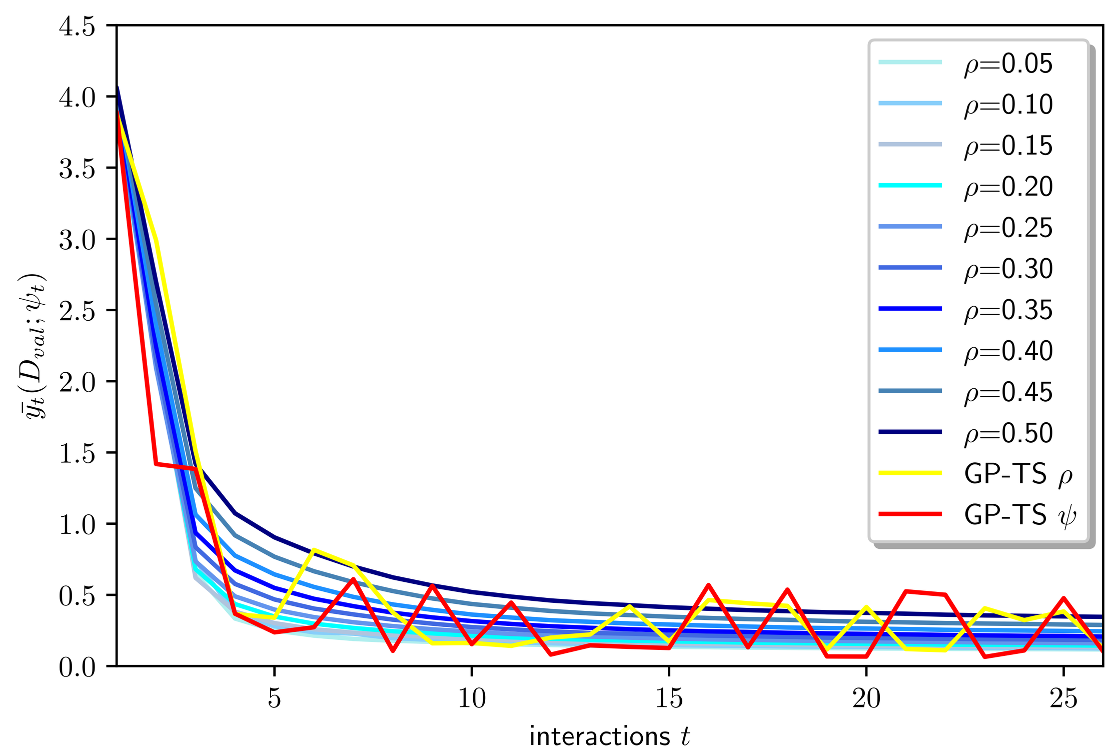
Averaged MLM validation performance comparison (lower is better) of grid-search based and the GP-TS based from *scratch* pre-trained RoBERTa models over interactions.



(a) wiki-c4.



(b) mimic.



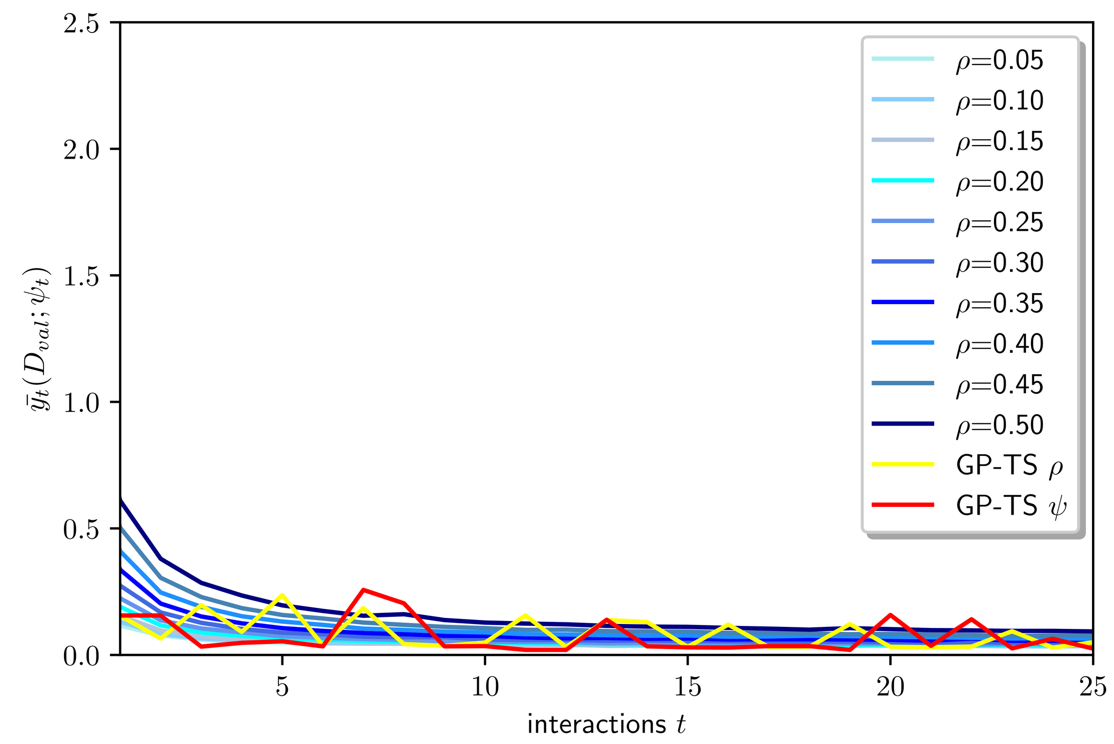
(c) e-commerce.

MLM dynamic masking experiments

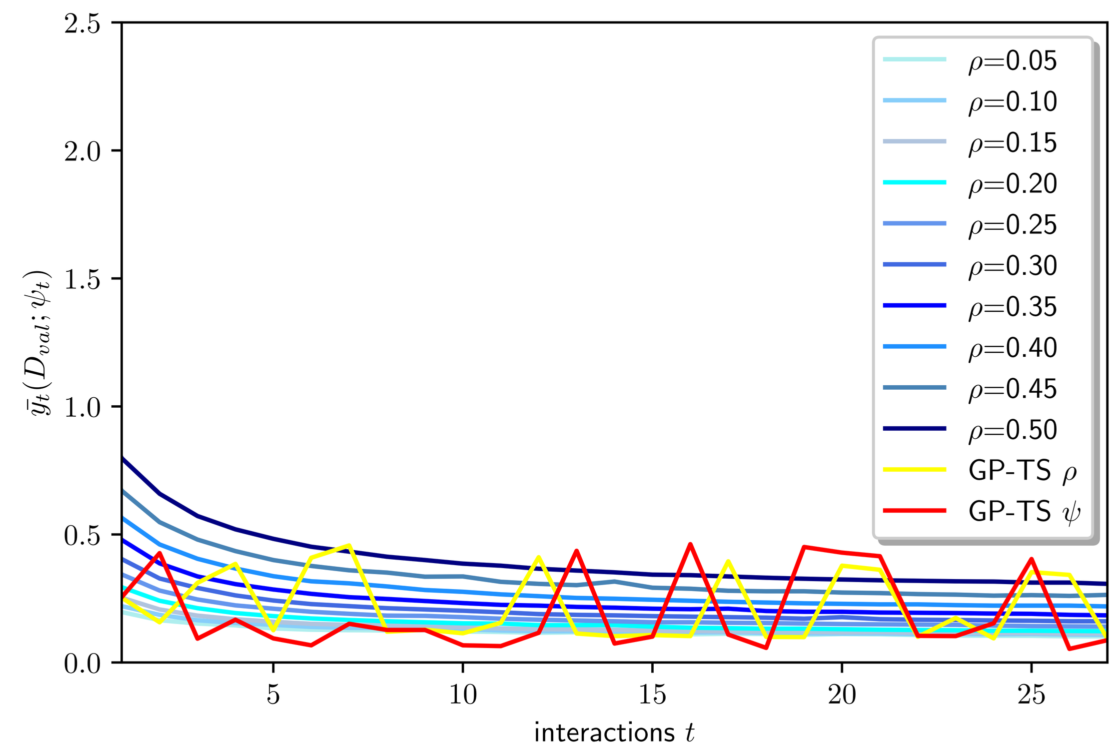
- **wiki-c4**: Wikitext-103 and Google's c4 RealNews datasets: average of 35 words per-sentence, more than 4,500M words total
- **mimic**: MIMIC-III Clinical database, with deidentified notes and reports for patients at intensive care unit: average of 200 words per-sentence, more than 400M words total
- **e-commerce**: A random subset of eBay marketplace product titles, descriptions and reviews: average of 5 words per-sentence, about 4,000M words total

Pre-training RoBERTa models *continually*

Averaged MLM validation loss performance comparison (lower is better) of grid-search based and the GP-TS based *continually* pre-trained RoBERTa models over interactions.



(a) mimic.



(b) e-commerce.

Pre-training experiments

- GP-TS pre-trains best RoBERTa models across datasets (scratch and continually)
- GP-TS MLM loss values fluctuate across interactions
- GP-TS pre-trains models with the lowest MLM, in less interactions.
- GP-TS selects sequences of hyperparameters, over a multi-dimensional space ψ

GP-TS

Superior and accelerated MLM dynamic masking pre-training performance

✓ Pre-training efficiency critical in practice
✓ Significant resource utilization savings (Grid-search can be avoided)

Follow-up work:

Downstream NLP task performance