Domain Adaptation for Robust Model Routing

Christoph Dann
Google Research
cdann@cdann.net

Yishay Mansour
Tel Aviv University and Google Research
mansour.yishay@gmail.com

Teodor V. Marinov
Google Research
tvmarinov@google.com

Mehryar Mohri
Google Research and Courant Institute of Mathematical Sciences, New York
mohri@google.com

Abstract

The rapid proliferation of domain-specialized machine learning models presents a challenge: while individual models excel in specific domains, their performance varies significantly across diverse applications. This makes selecting the optimal model for new tasks, especially with limited or no domain-specific data, a difficult problem. We address this challenge by formulating it as a multiple-source domain adaptation (MSA) problem. We introduce a novel, scalable algorithm that effectively routes each input to the best-suited model from a pool of available models. Our approach provides a key performance guarantee: for any new domain that lies within the convex hull of the source domains, the accuracy achieved by the best source model is maintained. This guarantee is formally established through a theoretical bound on the regret for new domains, expressed as a convex combination of the best regrets in the source domains, plus a concentration term that diminishes as the amount of source data increases.

1 Introduction

Fine-tuning is a key step in adapting large language models (LLMs) to specialized tasks or domains after their general pre-training. In this process, an LLM trained on vast datasets is further trained on smaller, task-specific datasets. As organizations and researchers fine-tune LLMs for tasks like summarization, translation, or customer service, the result is a growing collection of models, each optimized for different tasks but based on the same underlying architecture.

Routing algorithms are crucial for efficiently managing this diversity of specialized models, by determining which model best fits a given input. Recently, various routing algorithms have been proposed (Chen et al., 2023; Wang et al., 2023; Hu et al., 2024; Madaan et al., 2023; Yue et al., 2023; Lee et al., 2023; Shnitzer et al., 2023; Narayanan Hari and Thomson, 2023; Lu et al., 2023), including some with strong theoretical guarantees (Mao et al., 2023, 2024a,b). While these routing solutions can be effective for inputs drawn from each specific task distributions, they provide no guarantees for inputs drawn from a mixture of tasks. Building a fine-tuned model for every possible task combination is impractical, so how can routing be designed to handle such mixed-task inputs?

To address this problem, this paper frames model routing as a multiple-source domain adaptation (MSA) problem (Mansour et al., 2008) and derives a principled solution for enhancing robustness and adaptability across diverse and dynamic task distributions. Our approach grounded in strong MSA theory (Mansour et al., 2008, 2012; Hoffman et al., 2021; Cortes et al., 2021b) ensures that our routing model system performs as well as the best individual expert model across any task mixture. Furthermore, our solution is easily implemented and compatible with existing router training approaches. It enhances existing router training by strategically adjusting task domain weights.

Preprint. Under review.

Section 3 introduces our novel algorithm, which is supported by strong theoretical results (Section 4) and validated through extensive experimentation (Section 5). Related work in routing and multiple-source adaptation is reviewed in Appendix A. We begin by outlining our problem formulation.

2 Problem Formulation

We first introduce the model routing problem and then cast it as an MSA problem.

2.1 Model Routing

We consider a finite set of generative models, denoted by Π, where each model $\pi: \mathcal{X} \to \Delta(\mathcal{Y})$ maps inputs $\mathcal{X}$ to probability distributions over outputs $\mathcal{Y}$. For example, if Π consists of generative language models, $\mathcal{X}$ would represent prompts and $\mathcal{Y}$ their corresponding generations. Additionally, we assume there are k benchmark tasks, $D_1, \ldots, D_k$, where each D_i is a distribution over inputs. Typically, access to D_i is limited to a finite dataset. We will denote by $\hat{D}_i$ the empirical distribution consisting of n_i i.i.d. samples drawn from D_i. Let $r^\star: \mathcal{X} \times \mathcal{Y} \to [0,1]$ represent a scoring function that evaluates the quality of a generation $y \in \mathcal{Y}$ for a given input $x \in \mathcal{X}$. For example, $r^\star$ could indicate the probability that human evaluators prefer y over the output of a reference model. Although $r^\star$ may be unknown, we assume access to a scoring oracle $\mathcal{R}$ that provides unbiased estimates of $r^\star$ for any input-output pair (x, y). For simplicity, we assume that the scoring function $r^\star$ is uniform across all benchmark tasks, though this assumption can be relaxed. The *value* of a model $\pi \in \Pi$ on an input x or distribution over inputs D is defined as follows:

$$v(\pi, x) = \mathop{\mathbb{E}}_{y \sim \pi(x)}[r^\star(x, y)] \qquad v(\pi, D) = \mathop{\mathbb{E}}_{x \sim D}[v(\pi, x)].$$

Goal of predictive model routing. Given access to Π, $\mathcal{R}$, and the datasets $\widehat{D}_1, \ldots, \widehat{D}_k$, our goal is to select a high-quality probabilistic *routing function* $f: \mathcal{X} \to \Delta(\Pi)$ from a family $\mathcal{F}$ of such functions. Each routing function maps an input $x \in \mathcal{X}$ to a probability distribution over the models in Π. For any input x, a model $\pi \in \Pi$ is selected by sampling from the distribution $f(x)$.

For any $\pi: \mathcal{X} \to \Delta(\mathcal{Y})$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let $\pi(y|x)$ denote the probability of y under the distribution $\pi(x)$. Given a routing function $f \in \mathcal{F}$, we define the induced distribution $\pi_f(\cdot|x)$ over outputs $\mathcal{Y}$ as:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad \pi_f(y|x) = \sum_{\pi \in \Pi} f(\pi|x)\pi(y|x).$$

The objective is for f to route inputs x, drawn from an unknown test domain $D \in \Delta(\mathcal{X})$, to models in Π that yield high scores according to the oracle $\mathcal{R}$. Specifically, we aim to find an f that maximizes the expected score $v(\pi_f, D)$, without prior knowledge of D. The performance of a routing function f is evaluated by the *regret* of its induced policy π_f on the test domain D, defined as:

$$\mathrm{reg}(\pi_f, D) \coloneqq \max_{\pi \in \Pi} v(\pi, D) - v(\pi_f, D), \tag{1}$$

that is the gap between the performance of the best model in Π and that of the model selected by f.

Why is the test domain unknown? The test distribution D, representing the real-world data an application will encounter, is typically unknown during development. This is particularly true for new applications, where we lack sufficient data to accurately assess how the model will be used. Even for existing applications, the distribution D can change as user behavior evolves in response to model updates. For example, if a model demonstrates unexpected proficiency in a specific task, users might shift their usage patterns accordingly.

2.2 Predictive Model Routing as Multiple-Source Domain Adaptation.

Multiple-source domain adaptation (MSA) is a closely related problem that has been extensively studied, particularly in classification and regression problems (Mansour et al., 2008, 2012; Hoffman et al., 2021; Cortes et al., 2021b). In MSA, the task involves multiple source domains, $D_1, \ldots, D_k$, each associated with a near-optimal model $h_1, \ldots, h_k$ (Mansour et al., 2008). The target domain, D_λ, is defined as a λ-mixture of the source domains, $D_\lambda = \frac{1}{k}\sum_{i=1}^k \lambda_i D_i$, where $\lambda \in \Delta([k])$ represents unknown mixture weights. The objective is to devise a combination rule for the models h_i such that the resulting model performs well on any target domain D_λ.

We can formulate the predictive model routing problem as a multiple-source domain adaptation task by first selecting an appropriate model, π_i, for each dataset, which we refer to as the expert model for domain D_i. In many applications, natural choices for π_i arise, such as when a model π has been fine-tuned to perform well on a specific domain D_i. More generally, we can define π_i as the model in the set Π that achieves the highest value estimate for D_i. Next, we augment the empirical distributions $\widehat{D}_1, \ldots, \widehat{D}_k$ with score samples from each expert model. For each input x in the support of $\widehat{D}_i$, we compute scores $r_1, \ldots, r_k$ by generating responses $y_j \sim \pi_j(\cdot|x)$ from each expert π_j and querying the reward oracle, which returns scores $r_j \sim \mathcal{R}(x, y_j)$. These scores, r_j, serve as unbiased estimates of the value $v(\pi_j, x)$. We denote the augmented version of $\widehat{D}_i$ as $\bar{D}_i$.

With the *score-augmented distributions* $(\bar{D}_i)_{i\in[k]}$ in hand, the objective is to find a routing function (or combination rule) $f\colon \mathcal{X} \to \Delta([k])$ that maps inputs to a distribution over expert models. This routing function induces a mixed generation policy $\pi_f(y|x) = \sum_{i=1}^k f(i|x)\pi_i(y|x)$, which is evaluated based on its performance across any target domain D_λ. The quality of the routing function f is measured by its regret relative to the full policy set Π, as defined in (1). For the remainder of the paper, we adopt this domain adaptation perspective on predictive model routing, assuming that we are provided with a score-augmented empirical distribution $\bar{D}_i$ for each domain D_i and that the goal is to learn an effective routing function to the expert models.

3 Proposed Algorithm

To ensure robustness in model routing across test domains, we draw on two key areas of research: multiple-source domain adaptation (Mansour et al., 2008; Cortes et al., 2021b) and minimax-regret optimization (Alaiz-Rodrıguez et al., 2007; Rigter et al., 2021; Mohri et al., 2019; Agarwal and Zhang, 2022). Our approach is particularly aligned with the approaches of Cortes et al. (2021b) and Mohri et al. (2019); Agarwal and Zhang (2022). Specifically, we adopt the mixture over test domains and the associated theoretical guarantees from (Cortes et al., 2021b), while the objective formulation and optimization strategy are inspired by (Mohri et al., 2019; Agarwal and Zhang, 2022).

To design our algorithm, we begin by considering the idealized infinite-data setting and then introduce finite-sample approximations. Rather than minimizing regret under a fixed distribution, as defined in (1), we adopt a more robust objective inspired by the minimax regret optimization literature (Alaiz-Rodrıguez et al., 2007; Rigter et al., 2021; Mohri et al., 2019; Agarwal et al., 2017). Specifically, we aim to *minimize the worst-case regret over all possible test domains*:

$$\min_{f\in\mathcal{F}} \max_{\lambda\in\Delta([k])} \max_{\pi'\in\Pi} v(\pi', D_\lambda) - v(\pi_f, D_\lambda). \tag{2}$$

However, solving this optimization problem during training is challenging due to the maximization over $\pi' \in \Pi$. To address this challenge, we propose and explore two practical variants that avoid optimization over π'. Each variant minimizes regret relative to a specific policy, denoted as $\pi^\star_A$ or $\pi^\star_B$.

Option A: Pointwise Comparator. In this first variant, we aim to compete against a policy $\pi^\star_A$ that, for each input context x, achieves the performance of the best expert model. Formally, $v(\pi^\star_A, x) = \max_{i\in[k]} v(\pi_i, x)$ for all x. This leads to the following objective:

$$\min_{f\in\mathcal{F}} \max_{\lambda\in\Delta([k])} \mathcal{L}_A(f,\delta) \quad := \min_{f\in\mathcal{F}} \max_{\lambda\in\Delta([k])} v(\pi^\star_A, D_\lambda) - v(\pi_f, D_\lambda). \tag{3}$$

In the finite-sample setting, this min-max objective becomes:

$$\min_{f\in\mathcal{F}} \max_{\lambda\in\Delta([k])} \widehat{\mathcal{L}}_A(f,\delta) \quad := \min_{f\in\mathcal{F}} \max_{\lambda\in\Delta([k])} \mathop{\mathbb{E}}_{\substack{i\sim\lambda \\ (x,r_1,\ldots,r_k)\sim\bar{D}_i}} \left[\max_{j\in[k]} r_j - \sum_{l=1}^k f(l|x)\, r_l\right]. \tag{4}$$

where the maximum is taken over expert scores for each sample. While being easy to implement, this approach introduces additional bias when there is high variance in the expert scores for a given input.

Option B: Domain Comparator. To limit bias in the finite-sample objective, we leverage the structure of the model routing problem by using $\pi^\star_B$ as the comparator in the regret calculation. This policy, $\pi^\star_B\colon \mathcal{X} \times [k] \to \Delta(\mathcal{Y})$, takes both the input x and the domain label i, following the expert model π_i for samples from domain D_i; that is, $\pi^\star_B(x,i) = \pi_i(x)$. As we will demonstrate later, this fixed comparator provides strong regret guarantees without requiring an additional inner optimization over policies. This leads to the following optimization objective:

$$\min_{f \in \mathcal{F}} \max_{\lambda \in \Delta([k])} \mathcal{L}_B(f, \delta) \quad := \min_{f \in \mathcal{F}} \max_{\lambda \in \Delta([k])} v(\pi_B^\star, Q_\lambda) - v(\pi_f, D_\lambda) \tag{5}$$

with the finite-sample counterpart:

$$\min_{f \in \mathcal{F}} \max_{\lambda \in \Delta([k])} \widehat{\mathcal{L}}_B(f, \delta) \quad := \min_{f \in \mathcal{F}} \max_{\lambda \in \Delta([k])} \mathop{\mathbb{E}}_{\substack{i \sim \lambda \\ (x, r_1, \ldots, r_k) \sim \bar{D}_i}} \left[r_i - \sum_{l=1}^{k} f(l|x)\, r_l \right]. \tag{6}$$

Note that $\pi_A^\star$ and $\pi_B^\star$ coincide when domain experts are perfect, producing the best score for each individual x from their respective domain. However, in practice, even even π_i that are well-tuned for their domain D_i typically do not achieve this, which distinguishes $\pi_A^\star$ from $\pi_B^\star$ in general.

Algorithm. We follow the standard approach and tackle the saddle-point problems in Equation 4 or 6 as a two-player game, which can be solved by dueling two no-regret learners (see Mohri et al. (2019) for a general Mirror descent solution). Our algorithm is shown in Algorithm 1. The max-player can be solved efficiently with Hedge (Littlestone and Warmuth, 1994). For the min-player, we do not prescribe the exact update for f_t as we do not wish to prescribe a specific function class $\mathcal{F}$. Instead, we follow prior work (e.g. Cheng et al., 2022) and rely on an online learning oracle with we refer to as OLO. We assume that this oracle is a no-regret learner, which we formalize in Definition 1 in Appendix B. For finite context spaces, OLO can be instantiated as one Hedge instance per context with regret bound $O\left(\sqrt{kT|\mathcal{X}|\ln k}\right)$. In general, there is a large family of online-learning algorithms available with appropriate guarantees (Cesa-Bianchi and Lugosi, 2006).

Practical Implementation. Algorithm 1 can be seamlessly integrated into existing model training frameworks. For instance, in the case of language model routing, the class $\mathcal{F}$ can be a moderate-sized language model architecture, where the initial policy f_1 is a pre-trained model with its final layer replaced by a randomly initialized linear layer. At each round $t \in [T]$, a batch of samples is drawn from the augmented datasets, with equal proportions from each. The Hedge update of domain weights λ_t can be efficiently computed in closed form with minimal computational cost.

The update of f_t is handled using standard gradient-based optimizers on the objectives in (4) or (6), augmented with a KL-regularization, similar to RLHF training objectives (Christiano et al., 2017), such as regularization toward a uniform domain distribution or a given domain prior. Alternatively, the model can be optimized with a logistic proxy loss, similar to standard supervised fine-tuning objectives, which we explore further in Appendix B.3. Finally, Algorithm 1 returns an averaged model $\bar{f}$, where $\bar{f}(i|x) = \frac{1}{T}\sum_{t=1}^{T} f_t(i|x)$ for all $x \in \mathcal{X}$ and $i \in [k]$. While exact output averaging might not always be feasible, we can adopt a "model souping" approach by averaging the parameters θ_t of the models f_t across iterations. The final model is then represented by $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$, a technique that has proven effective in practice (Wortsman et al., 2022; Ramé et al., 2024).

4 Theoretical Guarantees

We derive performance guarantees for $\pi_{\bar{f}}$ returned by Algorithm 1 under both options and for different online learning oracles used for f_t updates in the appendix. We here present the following corollary for Option B as we find it most informative for the types of theoretical guarantees which we derive in the appendix.

Corollary 1. *Let $\mathcal{F}$ be a convex set. Then, with probability at least $1 - O(\delta)$ the regret of the function $\bar{f}$ returned by Algorithm 1 with Option B satisfies for all $\lambda \in \Delta(k)$ the following inequality:*

$$\mathrm{reg}(\pi_{\bar{f}}, D_\lambda) \le \mathrm{reg}(\pi_B^\star, D_\lambda) + O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log\left(\frac{|\mathcal{F}|}{\delta}\right)}{n_i}} + \max_i \frac{\lambda_i \log\left(\frac{|\mathcal{F}|}{\delta}\right)}{n_i} \right) + O\left(\frac{1}{\sqrt{T}}\right),$$

provided that $\mathcal{F}$ contains $f_{\lambda,\widehat{D}}$ for every $\lambda \in \Delta_k$, where $f_{\lambda,\widehat{D}}$ is defined as $f_\lambda(i|x) = \frac{\lambda_i \widehat{D}_i(x)}{\sum_{j=1}^{k}\lambda_j \widehat{D}_j(x)}$.

Recall that $\mathrm{reg}(\pi_B^\star, D_\lambda)$ is the regret of the policy, $\pi_B^\star$, which assigns any $x^{(i)} \sim D_i$ to its domain expert, π_i. Choosing λ as the i-th corner of the simplex, that is $\lambda_i = 1, \lambda_{j \ne i} = 0$, we see that

Algorithm 1: Domain adaptation for model routing algorithm

Input: Score-augmented distributions $\bar{D}_i$ for $i \in [k]$ of size n_i. Each sample is of the form $(x, r_1, \ldots, r_k)$ where x is the context and r_j is a reward estimate for expert policy π_j;

Output: Routing policy $f\colon \mathcal{X} \to \Delta_k$;

Initialize $\lambda_1 = [\frac{1}{k}, \ldots, \frac{1}{k}]^\top$ and f_1 in $\mathcal{F}$ arbitrarily;

for $t = 1, 2, \ldots, T$ **do**

  Sample $(x_t^{(i)}, r_{t,1}^{(i)}, \ldots, r_{t,k}^{(i)}) \sim \bar{D}_i$ for each $i \in [k]$;

  Determine benchmark scores with **option A** $c_t^{(i)} = \max_{j \in k} r_{t,j}^{(i)}$ or **option B** $c_t^{(i)} = r_{t,i}^{(i)}$;

  Max-player: Hedge

  Update $\lambda_{t+1} \propto \lambda_t \exp(-\gamma \ell_t)$ with losses $\ell_t : \ell_{t,i} = c_t^{(i)} - \sum_{j=1}^{k} r_{t,j}^{(i)} f_t(j|x_t^{(i)})$.

  Min-player: no-regret online learning update

  Update f_{t+1} with contexts $x_t^{(i)}$ and losses $\ell_t^{(i)} : \ell_{t,j}^{(i)} = \lambda_{t,i} (c_t^{(i)} - r_{t,j}^{(i)})$;

return $\bar{f} = \frac{1}{T} \sum_{t=1}^{T} f_t$

$\mathrm{reg}(\pi_B^\star, D_\lambda)$ is just the regret of the i-th domain expert on D_i and so we can bound $\mathrm{reg}(\pi_B^\star, D_\lambda)$ by the worst case regret of the domain experts on their respective domains. The term containing λ comes from relating the empirical game played only on n_i samples from each D_i to the population game over D_i. This term indicates that, in the worst case, we have to pay for the domain from which we observe the least amount of data. Finally, the term $O(1/\sqrt{T})$ comes from the regret of the OLO and concentration of other terms in the T-round empirical game solved by Algorithm 1. Overall, Corollary 1 shows that the regret of $\pi_{\bar{f}}$ is not much worse compared to the regret of the domain experts, up to concentration and terms related to solving the empirical game.

5 Empirical Evaluation

To demonstrate the effectiveness of Algorithm 1 in generating robust routing functions, we compare it against non-robust baselines on the MixInstruct benchmark by Jiang et al. (2023). This benchmark consists of 5 individual datasets. Each dataset corresponds to a domain $\widehat{D}_i$ and contains samples with prompts and various metrics for the generations of 11 open-source LLMs. For our analysis, we focus exclusively on the BLEU score and select the model with the highest average BLEU score per domain from the training split to serve as the domain expert π_i. The routing function f is initialized using a pre-trained Gemma 2B model (Team et al., 2024), with the final layer replaced by a fully connected, randomly initialized layer.

Several prior studies have explored optimal strategies for learning a routing function tailored to specific data distributions (Jiang et al., 2023; Hu et al., 2024)—among others. We view our algorithm as a framework that can enhance these approaches through the OLO oracle. Thus, our experiments aim not to compare different learning methodologies but to assess the impact of robust routing by adjusting the domain weights during training. Specifically, we compare Algorithm 1 with and without updates to λ_t (i.e., $\gamma = 0$ vs. $\gamma \neq 0$), while keeping all other parameters constant.

		regret vs best expert		regret vs domain expert	
Loss for f	Option	Baseline	Alg 1	Baseline	Alg 1
linear	A	4.60	**4.28**	1.65	**0.49**
linear	B	**4.60**	7.09	1.64	**1.08**
log	A	2.70	**2.37**	-0.06	**-0.39**
log	B	7.90	**7.84**	0.58	**0.23**

Table 1: Overview of regret in the worst-case test domain comparing the routing function produced by Algorithm 1 against a routing function produced by training with uniform and fixed domain weights. Results are averages across 5 seeds. Algorithm 1 consistently reduces the regret against the competitor targeted by the selected option.

6 Conclusion

We presented a novel approach for combining multiple domain expert algorithms using online learning oracles, achieving regret bounds that are tied to the performance of these oracles. Our method leverages theoretical guarantees, ensuring robustness in a variety of settings. Additionally, we validated the effectiveness of our approach through experiments on the MixInstruct dataset, where the results highlight the practical benefits of our model routing strategy.

References

A. Agarwal and T. Zhang. Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory*, pages 2704–2729. PMLR, 2022.

A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

R. Alaiz-Rodrıguez, A. Guerrero-Curieses, and J. Cid-Sueiro. Minimax regret classifier for imprecise class distributions. *Journal of Machine Learning Research*, 8:103–130, 2007.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

L. Chen, M. Zaharia, and J. Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

C.-A. Cheng, T. Xie, N. Jiang, and A. Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pages 3852–3878. PMLR, 2022.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

C. Cortes, M. Mohri, and A. T. Suresh. Relative deviation margin bounds. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2122–2131. PMLR, 2021a.

C. Cortes, M. Mohri, A. T. Suresh, and N. Zhang. A discriminative technique for multiple-source adaptation. In *International Conference on Machine Learning*, pages 2132–2143. PMLR, 2021b.

G. DeSalvo, M. Mohri, and U. Syed. Learning with deep cascades. In *Algorithmic Learning Theory: 26th International Conference, ALT 2015, Banff, AB, Canada, October 4-6, 2015, Proceedings 26*, pages 254–269. Springer, 2015.

D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: The importance of being improper. In *Conference on learning theory*, pages 167–208. PMLR, 2018.

E. Hazan, T. Koren, and K. Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209. PMLR, 2014.

J. Hoffman, M. Mohri, and N. Zhang. Multiple-source adaptation theory and algorithms. *Ann. Math. Artif. Intell.*, 89(3-4):237–270, 2021.

Q. J. Hu, J. Bieker, X. Li, N. Jiang, B. Keigwin, G. Ranganath, K. Keutzer, and S. K. Upadhyay. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*, 2024.

D. Jiang, X. Ren, and B. Y. Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*, 2023.

S. M. Kakade and A. Ng. Online bounds for bayesian algorithms. *Advances in neural information processing systems*, 17, 2004.

C.-H. Lee, H. Cheng, and M. Ostendorf. Orchestrallm: Efficient orchestration of language models for dialogue state tracking. *arXiv preprint arXiv:2311.09758*, 2023.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

K. Lu, H. Yuan, R. Lin, J. Lin, Z. Yuan, C. Zhou, and J. Zhou. Routing to the expert: Efficient reward-guided ensemble of large language models. *arXiv preprint arXiv:2311.08692*, 2023.

A. Madaan, P. Aggarwal, A. Anand, S. P. Potharaju, S. Mishra, P. Zhou, A. Gupta, D. Rajagopal, K. Kappaganthu, Y. Yang, et al. Automix: Automatically mixing language models. *arXiv preprint arXiv:2310.12963*, 2023.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21, 2008.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the rényi divergence. *arXiv preprint arXiv:1205.2628*, 2012.

A. Mao, C. Mohri, M. Mohri, and Y. Zhong. Two-stage learning to defer with multiple experts. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

A. Mao, M. Mohri, and Y. Zhong. Principled approaches for learning to defer with multiple experts. In R. P. Barneva, V. E. Brimkov, C. Gentile, and A. Pacchiano, editors, *Artificial Intelligence and Image Analysis - 18th International Symposium on Artificial Intelligence and Mathematics, ISAIM 2024, and 22nd International Workshop on Combinatorial Image Analysis, IWCIA 2024, Fort Lauderdale, FL, USA, January 8-10, 2024, Revised Selected Papers*, volume 14494 of *Lecture Notes in Computer Science*, pages 107–135. Springer, 2024a.

A. Mao, M. Mohri, and Y. Zhong. Regression with multi-expert deferral. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024b.

H. B. McMahan and M. Streeter. Open problem: Better bounds for online logistic regression. In *Conference on Learning Theory*, pages 44–1. JMLR Workshop and Conference Proceedings, 2012.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 2019.

S. Narayanan Hari and M. Thomson. Tryage: Real-time, intelligent routing of user prompts to large language models. *arXiv e-prints*, pages arXiv–2308, 2023.

A. Ramé, J. Ferret, N. Vieillard, R. Dadashi, L. Hussenot, P.-L. Cedoz, P. G. Sessa, S. Girgin, A. Douillard, and O. Bachem. Warp: On the benefits of weight averaged rewarded policies. *arXiv preprint arXiv:2406.16768*, 2024.

M. Rigter, B. Lacerda, and N. Hawes. Minimax regret optimisation for robust planning in uncertain markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 (13), pages 11930–11938, 2021.

G. I. Shamir. Logistic regression regret: What's the catch? In *Conference on Learning Theory*, pages 3296–3319. PMLR, 2020.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

T. Shnitzer, A. Ou, M. Silva, K. Soule, Y. Sun, J. Solomon, N. Thompson, and M. Yurochkin. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*, 2023.

G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Y. Wang, K. Chen, H. Tan, and K. Guo. Tabi: An efficient multi-level inference system for large language models. In *Proceedings of the Eighteenth European Conference on Computer Systems*, pages 233–248, 2023.

M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

M. Yue, J. Zhao, M. Zhang, L. Du, and Z. Yao. Large language model cascades with mixture of thoughts representations for cost-efficient reasoning. *arXiv preprint arXiv:2310.03094*, 2023.

Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35: 7103–7114, 2022.

Contents of Appendix

A Related Work

The multiple-source adaptation (MSA) problem was theoretically studied by Mansour et al. (2008, 2012). Later, Hoffman et al. (2021) introduced an efficient algorithm based on domain density estimation. This approach was subsequently improved by Cortes et al. (2021b), who replaced density estimation with a domain classifier. However, despite this simplification, their method still requires solving a difference of convex (DC) programming problem, which may not be well-suited for modern LLM inference scenarios.

More recently, various types of routing problems in LLMs have been investigated. Post-hoc routing (Chen et al., 2023; Wang et al., 2023; Hu et al., 2024; Madaan et al., 2023; Yue et al., 2023; Lee et al., 2023) involves processing inputs with multiple expert LLMs and selecting the best output based on a scoring rule. A specific form of post-hoc routing, known as cascading routing, was studied by Chen et al. (2023); Wang et al. (2023); Yue et al. (2023); Hu et al. (2024), where inputs are processed sequentially by experts until a sufficiently high-quality response is obtained. Theoretical investigations of cascading ideas in classification have been conducted by DeSalvo et al. (2015).

Predictive routing (Shnitzer et al., 2023; Narayanan Hari and Thomson, 2023; Lu et al., 2023) offers an alternative, where an input is directed to a single expert LLM, which alone processes it. Mixture of Experts (MoEs) (Shazeer et al., 2017; Zhou et al., 2022) can also be seen as a form of predictive routing, where only a subset of an LLM's parameters is activated for processing each token. Mao et al. (2023, 2024a,b) have introduced deferral algorithms, which can be used in particular for routing applications, together with an extensive theoretical guarantees. Recent efforts by Hu et al. (2024) and Jiang et al. (2023) have proposed benchmarks for evaluating mixtures of LLMs. For a more comprehensive review of this literature, we refer readers to Hu et al. (2024). Our work focuses exclusively on the predictive routing setting.

B Theoretical Analysis

We first provide a definition that formalizes the notion of online-learning we assume for the updates of f:

Definition 1 (Online learning oracle)**.** *An algorithm* OLO *is referred to as an online learning oracle for a class* $\mathcal{F} \subseteq \mathcal{X} \to \Delta_k$ *if it satisfies the following condition. Given an arbitrary, potentially adversarial sequence of context-loss pairs* $(x_1, \ell_1, \ldots, x_T, \ell_T)$*,* OLO *observes each context* x_t *sequentially and maintains a sequence of policies* $f_{t+1} \in \mathcal{F}$*, updating the policy after observing each loss sample* ℓ_t*. The regret of* OLO *is given by:*

$$\mathrm{Reg}_{\mathcal{F}}(T) = \max_{f \in \mathcal{F}} \sum_{t=1}^{T} \langle f(x_t) - f_t(x_t), \ell_t \rangle = o(T),$$

and is sublinear with probability at least $1 - \delta$*.*

We note that in Algorithm 1 the losses required by Definition 1, ℓ_t, for each round $t \in [T]$ are the sum over the per-domain losses, that is $\ell_t = \sum_{i=1}^{k} \ell_t^{(i)} = \sum_{i=1}^{k} \lambda_{t,i}(c_t^{(i)} - r_{t,j}^{(i)})$.

Using an OLO we show the following regret guarantee for Algorithm 1.

Theorem 1. *Let* $\mathcal{F}$ *be a convex set. Then, with probability at least* $1 - O(\delta)$*, the regret of the function* $\bar{f}$ *returned by Algorithm 1 with Option A satisfies for all* $\lambda \in \Delta(k)$ *the following inequality:*

$$\mathrm{reg}(\pi_{\bar{f}}, D_\lambda) \le \mathrm{reg}(\pi_A^\star, D_\lambda) + \widehat{V}_A^\star + \frac{\mathrm{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log k + C_\delta}{T}}\right) + O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 C_\delta}{n_i}} + \max_i \frac{\lambda_i C_\delta}{n_i}\right),$$

where $\widehat{V}_A^\star = \max_{\lambda \in \Delta_k} \inf_{f \in \mathcal{F}} \widehat{\mathcal{L}}(f, \lambda)$ *is the optimal value of the objective in Equation 4,* $\mathrm{Reg}_{\mathcal{F}}(T) = o(T)$ *is the regret of the* OLO *oracle,* $C_\delta = \log(\frac{|\mathcal{F}|}{\delta})$*, and* $\pi_A^\star$ *is the competitor policy for option A. The same guarantee holds for Option B with* $\pi_A^\star$ *is replaced by* $\pi_B^\star$ *and* $\widehat{V}_A^\star$ *by* $\widehat{V}_B^\star$*.*

The performance guarantee for both options is the same up to the first two terms. By construction, $\pi_A^\star$ is a stronger competitor than $\pi_B^\star$, since the inequality $v(\pi_B^\star, x) \le \max_{i \in [k]} v(\pi_i, x) = v(\pi_A^\star, x)$

holds for all $x \in \mathcal{X}$. Option A may therefore seem preferable as the first term in the guarantee is always more favorable than for B. However, we expect that in most cases $\widehat{V}_B^\star \le \widehat{V}_A^\star$ since $V_B^\star$ is small under much weaker conditions than $\widehat{V}_A^\star$.

We prove Theorem 1 by deriving two separate guarantees for the different options in Algorithm 1 in Theorem 2 for Option B and in Theorem 3 for Option A. Recall the definitions of the objectives used by our algorithms as

$$\mathcal{L}_A(f,\lambda) = \mathop{\mathbb{E}}_{i\sim\lambda} \mathop{\mathbb{E}}_{x\sim D_i} \mathop{\mathbb{E}}_{j\sim f(x)} \left[\max_m v(\pi_m, x) - v(\pi_j, x)\right] \tag{7}$$

$$\widehat{\mathcal{L}}_A(f,\lambda) = \mathop{\mathbb{E}}_{i\sim\lambda} \mathop{\mathbb{E}}_{(x,r_1,\dots,r_k)\sim\widehat{D}_i} \mathop{\mathbb{E}}_{j\sim f(x)} \left[\max_m r_m - r_j\right] \tag{8}$$

$$\mathcal{L}_B(f,\lambda) = \mathop{\mathbb{E}}_{i\sim\lambda} \mathop{\mathbb{E}}_{x\sim D_i} \mathop{\mathbb{E}}_{j\sim f(x)} \left[v(\pi_i, x) - v(\pi_j, x)\right] \tag{9}$$

$$\widehat{\mathcal{L}}_B(f,\lambda) = \mathop{\mathbb{E}}_{i\sim\lambda} \mathop{\mathbb{E}}_{(x,r_1,\dots,r_k)\sim\widehat{D}_i} \mathop{\mathbb{E}}_{j\sim f(x)} \left[r_i - r_j\right]. \tag{10}$$

In the following, we refer by $\mathcal{L}$ jointly to $\mathcal{L}_A$ or $\mathcal{L}_B$ and $\widehat{\mathcal{L}}$ to $\widehat{\mathcal{L}}_A$ or $\widehat{\mathcal{L}}_B$ respectively.

Lemma 1. *The objectives $\mathcal{L}_A, \mathcal{L}_B, \widehat{\mathcal{L}}_A, \widehat{\mathcal{L}}_B$ are bilinear in f and λ. If $\mathcal{F} \subseteq \mathcal{X} \to \Delta_k$ is convex, then*

$$\inf_{f\in\mathcal{F}} \max_{\lambda\in\Delta_k} \mathcal{L}_A(f,\lambda) = \max_{\lambda\in\Delta_k} \inf_{f\in\mathcal{F}} \mathcal{L}_A(f,\lambda). \tag{11}$$

and analogously for $\mathcal{L}_B, \widehat{\mathcal{L}}_A$ and $\widehat{\mathcal{L}}_B$.

Proof. We see directly from (7) that all objectives are linear in both arguments. The second part follows from Sion's minimax theorem, since both Δ_k and $\mathcal{F}$ are convex and Δ_k is compact. □

The following lemma shows that the costs and rewards concentrate around their expectations.

Lemma 2. *The following hold*

$$\mathbb{P}\left(\sup_{\lambda\in\Delta(k)} \sum_{t=1}^T \sum_{i=1}^k \lambda_i \left(c_t^{(i)} - \mathbb{E}[c_t^{(i)}] - \sum_{j=1}^k f_t(j|x_t^{(i)})(r_{t,j}^{(i)} - \mathbb{E}[r_{t,j}^{(i)}])\right) \ge 2\sqrt{T\log(k/\delta)}\right) \le \delta$$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \sum_{t=1}^T \sum_{i=1}^k \lambda_{t,i} \left(c_t^{(i)} - \mathbb{E}[c_t^{(i)}] - \sum_{j=1}^k f(j|x_t^{(i)})(r_{t,j}^{(i)} - \mathbb{E}[r_{t,j}^{(i)}])\right) \ge 2\sqrt{T\log(|\mathcal{F}|/\delta)}\right) \le \delta.$$

Proof. We start by showing the first inequality. First note that for every $i \in [k]$, $\{c_t^{(i)} - \mathbb{E}[c_t^{(i)}] - \sum_{j=1}^k f_t(j|x_t^{(i)})(r_{t,j}^{(i)} - \mathbb{E}[r_{t,j}^{(i)}])\}_{t\in[T]}$ is a martingale difference sequence with respect to the filtration created by the online oracle. Azuma-Hoeffding's inequality and a union bound implies that

$$\mathbb{P}\left(\sup_{i\in[k]} \sum_{t=1}^T \left(c_t^{(i)} - \mathbb{E}[c_t^{(i)}] - \sum_{j=1}^k f_t(j|x_t^{(i)})(r_{t,j}^{(i)} - \mathbb{E}[r_{t,j}^{(i)}])\right) \ge 2\sqrt{T\log(k/\delta)}\right) \le \delta.$$

Next, we have

$$\begin{aligned}
&\sup_{\lambda\in\Delta(k)} \sum_{i=1}^k \lambda_i \sum_{t=1}^T \left(c_t^{(i)} - \mathbb{E}[c_t^{(i)}] - \sum_{j=1}^k f_t(j|x_t^{(i)})(r_{t,j}^{(i)} - \mathbb{E}[r_{t,j}^{(i)}])\right) \\
&= \sup_{i\in[k]} \sum_{t=1}^T \left(\mathbb{E}[c_t^{(i)}] - c_t^{(i)} - \sum_{j=1}^k f_t(j|x_t^{(i)})(\mathbb{E}[r_{t,j}^{(i)}] - r_{t,j}^{(i)})\right),
\end{aligned}$$

since $\sum_{i=1}^k \lambda_i \sum_{t=1}^T \left(c_t^{(i)} - \mathbb{E}[c_t^{(i)}] - \sum_{j=1}^k f_t(j|x_t^{(i)})(r_{t,j}^{(i)} - \mathbb{E}[r_{t,j}^{(i)}])\right)$ is linear in λ and the supremum will be achieved at one of the corners of the probability simplex.

The second inequality holds in a similar way by using Azuma-Hoeffding's inequality and a union bound over $\mathcal{F}$. □

We note that the notation $\log(|\mathcal{F}|)$ is overloaded to mean the metric entropy for function classes which have infinite cardinality. For the rest of the paper we consider $\log(|\mathcal{F}|)$ to be the metric entropy with respect to the following distance $d(f, f') = \sup_{x \in \mathcal{X}} \|f(x) - f'(x)\|_1$.

Lemma 3. *Let $\bar{f} = \frac{1}{T}\sum_{t=1}^{T} f_t, \bar{\lambda} = \frac{1}{T}\sum_{t=1}^{T} \lambda_t$ be the average iterates of Algorithm 1. Then*

$$\max_{\lambda \in \Delta_k, f \in \mathcal{F}} [\widehat{\mathcal{L}}(\bar{f}, \lambda) - \widehat{\mathcal{L}}(f, \bar{\lambda})] \le \frac{\mathrm{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right) \tag{12}$$

with high probability at least $1 - O(\delta)$, where $\mathrm{Reg}_{\mathcal{F}}(T)$ is the regret of the online learning oracle from Definition 1.

Proof. We begin by noting that

$$\widehat{\mathcal{L}}(\bar{f}, \lambda) = \mathbb{E}_{i \sim \lambda, x^{(i)} \sim \bar{D}_i}\left[\sum_{i=1}^{k} \lambda_i (c^{(i)} - \langle \bar{f}, r^{(i)} \rangle\right] = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} \lambda_i \left(\mathbb{E}[c_t^{(i)}] - \sum_{j=1}^{k} f_t(j|x_t^{(i)})\,\mathbb{E}[r_{t,j}^{(i)}]\right)$$

$$\widehat{\mathcal{L}}(f, \bar{\lambda}) = \mathbb{E}_{i \sim \lambda, x^{(i)} \sim \bar{D}_i}\left[\sum_{i=1}^{k} \bar{\lambda}_i (c^{(i)} - \langle f, r^{(i)} \rangle\right] = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} \lambda_{t,i} \left(\mathbb{E}[c_t^{(i)}] - \sum_{j=1}^{k} f(j|x_t^{(i)})\,\mathbb{E}[r_{t,j}^{(i)}]\right)$$

Further, using Lemma 2 we have that w.p. $1 - \delta$ for all $f \in \mathcal{F}$ and all $\lambda \in \Delta(k)$

$$\begin{aligned}
&\widehat{\mathcal{L}}(\bar{f}, \lambda) - \widehat{\mathcal{L}}(f, \bar{\lambda}) \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} \lambda_i \left(\mathbb{E}[c_t^{(i)}] - \sum_{j=1}^{k} f_t(j|x_t^{(i)})\,\mathbb{E}[r_{t,j}^{(i)}]\right) - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} \lambda_{t,i} \left(\mathbb{E}[c_t^{(i)}] - \sum_{j=1}^{k} f(j|x_t^{(i)})\,\mathbb{E}[r_{t,j}^{(i)}]\right) \\
&\le \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} \lambda_i \left(c_t^{(i)} - \sum_{j=1}^{k} f_t(j|x_t^{(i)}) r_{t,j}^{(i)}\right) - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} \lambda_{t,i} \left(c_t^{(i)} - \sum_{j=1}^{k} f(j|x_t^{(i)}) r_{t,j}^{(i)}\right) + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \langle \lambda - \lambda_t, \ell_t' \rangle + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} \langle \ell_t^{(i)}, f_t(x_t^{(i)}) - f(x_t^{(i)}) \rangle + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right) \\
&\le \frac{\mathrm{Reg}_{\Lambda}(T)}{T} + \frac{\mathrm{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right).
\end{aligned}$$

Since $\mathrm{Reg}_{\Lambda}(T) = O\left(\sqrt{T\log(k|\mathcal{F}|/\delta)}\right)$ with probability at least $1 - O(\delta)$ the result follows. ∎

Lemma 4. *Let $V^\star = \inf_{f \in \mathcal{F}} \max_{\lambda \in \Delta_k} \mathcal{L}(f, \lambda)$ be the optimal value of the saddle-point. Then Algorithm 1 converges to that value with high probability at least $1 - O(\delta)$, that is,*

$$\max_{\lambda \in \Delta_k} \widehat{\mathcal{L}}(\bar{f}, \lambda) \le \widehat{V}^\star + \frac{\mathrm{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right).$$

This statement is true for either option A and option B.

Proof. By Lemma 1, the following chain of inequalities holds

$$\inf_{f \in \mathcal{F}} \widehat{\mathcal{L}}(f, \bar{\lambda}) \le \max_{\lambda \in \Delta_k} \inf_{f \in \mathcal{F}} \widehat{\mathcal{L}}(f, \lambda) = \widehat{V}^\star = \inf_{f \in \mathcal{F}} \max_{\lambda \in \Delta_k} \widehat{\mathcal{L}}(f, \lambda) \le \max_{\lambda \in \Delta_k} \widehat{\mathcal{L}}(\bar{f}, \lambda).$$

Rearranging terms yields

$$\begin{aligned}
\widehat{\mathcal{L}}(\bar{f}, \lambda) &\le \widehat{V}^\star + \max_{\lambda \in \Delta_k} \widehat{\mathcal{L}}(\bar{f}, \lambda) - \inf_{f \in \mathcal{F}} \widehat{\mathcal{L}}(f, \bar{\lambda}) \\
&\le \widehat{V}^\star + \frac{\mathrm{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right). && \text{(Lemma 3)}
\end{aligned}$$

∎

B.1 Analysis for Option B

Lemma 5 (Concentration for option B)**.** *For a fixed λ and $f \in \mathcal{F}$, we have with probability at least $1 - O(\delta)$*

$$\left|\mathcal{L}_B(f,\lambda) - \widehat{\mathcal{L}}_B(f,\lambda)\right| \le O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log(1/\delta)}{n_i}} + \max_i \frac{\lambda_i \log(1/\delta)}{n_i}\right)$$

and

$$\left|\mathcal{L}_B(f,\lambda) - \widehat{\mathcal{L}}_B(f,\lambda)\right| \le O\left(\sum_{i=1}^{k} \lambda_i \sqrt{\frac{\log(1/\delta)}{n_i}}\right)$$

Proof. Consider a fixed λ, f and $i \in [k]$. Order $\bar{D}_i$ arbitrarily and denote $(x_t, r_{t,1}, \dots, r_{t,k})$ the t-th datapoint in $\bar{D}_i$. Then $Y_{i,t} = \mathbb{E}_{j\sim f(x_t)}[r_{t,i} - r_{t,j}]$ are i.i.d. random variables with mean $\mathbb{E}\, Y_{i,t} = v(\pi_i, D_i) - v(\pi_f, D_i)$. Since scores are bounded, $Y_{i,t}$ centered to its mean is sub-Gaussian and we can bound with probability at least $1 - \delta$

$$\begin{aligned}
\mathcal{L}_B(f,\lambda) - \widehat{\mathcal{L}}_B(f,\lambda) &= \sum_{i=1}^{k} \frac{\lambda_i}{n_i} \sum_{t=1}^{n_i} [\mathbb{E}\, Y_{i,t} - Y_{i,t}] \\
&\le O\left(\sqrt{\sum_{i=1}^{k}\sum_{t=1}^{n_i} \frac{\lambda_i^2}{n_i^2} \log(1/\delta)} + \max_i \frac{\lambda_i \log(1/\delta)}{n_i}\right) \\
&= O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log(1/\delta)}{n_i}} + \max_i \frac{\lambda_i \log(1/\delta)}{n_i}\right)
\end{aligned}$$

$\square$

Lemma 6 (Value of the game for option B)**.** *Let $V_B^\star = \inf_{f\in\mathcal{F}} \max_{\lambda\in\Delta_k} \mathcal{L}_B(f,\lambda)$ be the optimal value of the saddle-point. Assume that the function class $\mathcal{F}$ contains f_λ for every $\lambda \in \Delta_k$, where $f_{\lambda,D}$ is defined as $f_\lambda(i|x) = \frac{\lambda_i D_i(x)}{\sum_{j=1}^k \lambda_j D_j(x)}$. Then the value of the game is non-positive, i.e., $V_B^\star \le 0$.*

Proof. Let $\lambda \in \Delta_k$ be arbitrary and consider $f(i|x) = \frac{\lambda_i D_i(x)}{\sum_{j=1}^k \lambda_j D_j(x)}$. We then have

$$\begin{aligned}
\mathcal{L}_B(f,\lambda) &= v(\pi_{dom}, Q_\lambda) - v(\pi_f, D_\lambda) \\
&= \sum_{i=1}^{k} \lambda_i \sum_{x\in\mathcal{X}} D_i(x) \langle \pi_{dom}(x,i), r^\star(x)\rangle - \sum_{x\in\mathcal{X}} D_\lambda(x)\langle \pi_f(x), r^\star(x)\rangle \\
&= \sum_{i=1}^{k} \lambda_i \sum_{x\in\mathcal{X}} D_i(x) \langle \pi_{dom}(x,i), r^\star(x)\rangle - \sum_{x\in\mathcal{X}} \sum_{i=1}^{k} \lambda_i D_i(x)\langle \pi_i(x), r^\star(x)\rangle \\
&\qquad\qquad \text{(definition of } f\text{)} \\
&= \sum_{i=1}^{k} \lambda_i \sum_{x\in\mathcal{X}} D_i(x) \langle \pi_{dom}(x,i) - \pi_i(x), r^\star(x)\rangle \\
&= 0 \qquad\qquad (\pi_{dom}(x,i) = \pi_i(x))
\end{aligned}$$

$\square$

Theorem 2 (Regret bound for Option B)**.** *Assume that the function class $\mathcal{F}$ is convex. Then the solution $\bar{f}$ returned by Algorithm 1 with Option B satisfies with probability at least $1 - O(\delta)$ for any fixed λ*

$$\operatorname{reg}(\pi_{\bar{f}}, D_\lambda) \le \sum_{i=1}^{k} \lambda_i \operatorname{reg}(\pi_i, D_i) + \widehat{V}_B^\star + \frac{\mathrm{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right) \tag{13}$$

$$+O\left(\sqrt{\sum_{i=1}^{k}\frac{\lambda_i}{n_i}\log\frac{|\mathcal{F}|}{\delta}}+\max_i\frac{\lambda_i}{n_i}\log\frac{|\mathcal{F}|}{\delta}\right) \tag{14}$$

Further, if the function class $\mathcal{F}$ contains $f_{\lambda,\widehat{D}}$ for every $\lambda \in \Delta_k$, where $f_{\lambda,\widehat{D}}$ is defined as $f_\lambda(i|x) = \frac{\lambda_i \widehat{D}_i(x)}{\sum_{j=1}^k \lambda_j \widehat{D}_j(x)}$, then $\widehat{V}_B^\star \le 0$. If this only holds on a population level, i.e., $\mathcal{F} \le \{f_{\lambda,D} : \lambda \in \Delta_k\}$, then we can still bound $\widehat{V}_B^\star = O\left(\sqrt{\frac{k\log(1/\delta)}{\min_i n_i}}\right)$.

Proof. We can decompose the regret of $\bar{f}$ on D_λ as

$$\begin{aligned}
\operatorname{reg}(\pi_{\bar{f}}, D_\lambda) &= \max_{\pi\in\Pi} v(\pi, D_\lambda) - v(\pi^\star, Q_\lambda) + v(\pi^\star, Q_\lambda) - v(\pi_{\bar{f}}, D_\lambda)\\
&= \max_{\pi\in\Pi} v(\pi, D_\lambda) - v(\pi^\star, Q_\lambda) + \mathcal{L}_B(\bar{f}, \lambda)\\
&\le \max_{\pi\in\Pi} v(\pi, D_\lambda) - v(\pi^\star, Q_\lambda) + \widehat{\mathcal{L}}_B(\bar{f}, \lambda) + O\left(\sqrt{\sum_{i=1}^{k}\frac{\lambda_i}{n_i}\log\frac{|\mathcal{F}|}{\delta}}+\max_i\frac{\lambda_i}{n_i}\log\frac{|\mathcal{F}|}{\delta}\right)
\end{aligned}$$

(Lemma 5)

where the last inequality follows from a union bound over $f \in \mathcal{F}$ and holds with probability at least $1 - O(\delta)$. The first two terms can be upper-bounded by the regret of each expert policy π_i on its own dataset, weighted by λ, i.e.,

$$\max_{\pi\in\Pi} v(\pi, D_\lambda) - v(\pi^\star, Q_\lambda) = \max_{\pi\in\Pi}\sum_{i=1}^{k}\lambda_i\left(v(\pi, D_i) - v(\pi_i, D_i)\right) \le \sum_{i=1}^{k}\lambda_i \operatorname{reg}(\pi_i, D_i).$$

We now bound $\widehat{\mathcal{L}}_B(\bar{f}, \lambda)$ further by Lemma 4 with probability at least $1 - O(\delta)$ as

$$\widehat{\mathcal{L}}_B(\bar{f}, \lambda) \le \widehat{V}_B^\star + \frac{\operatorname{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right).$$

Plugging both bounds in the previous decomposition yields the desired bound. For the bound on $\bar{V}_B^\star$, we apply Lemma 6 on $\widehat{D}$ directly or on D and apply Lemma 5 with a union bound over Δ_k. □

B.2 Analysis for Option A

Lemma 7 (Concentration for option A)**.** *For a fixed λ and $f \in \mathcal{F}$, we have with probability at least $1 - \delta$*

$$\begin{aligned}
\mathcal{L}_A(f,\lambda) - \widehat{\mathcal{L}}_A(f,\lambda) \le& O\left(\sqrt{\sum_{i=1}^{k}\frac{\lambda_i^2\log(1/\delta)}{n_i}}+\max_i\frac{\lambda_i\log(1/\delta)}{n_i}\right)\\
\widehat{\mathcal{L}}_A(f,\lambda) - \mathcal{L}_A(f,\lambda) \le& O\left(\sqrt{\sum_{i=1}^{k}\frac{\lambda_i^2\log(1/\delta)}{n_i}}+\max_i\frac{\lambda_i\log(1/\delta)}{n_i}\right) + \left|\sum_{i=1}^{k}\lambda_i \operatorname{bias}_A(i)\right|
\end{aligned}$$

where

$$\operatorname{bias}_A(i) = \frac{1}{n_i}\sum_{t=1}^{n_i}\left[\max_m v(\pi_m, x_t^{(i)}) - \mathop{\mathbb{E}}_{r_1,\dots,r_m|x=x_t^{(i)}}[\max_m r_m]\right].$$

Proof. Consider and ordering of the samples in each augmented dataset and denote by $(x_t^{(i)}, r_{t,1}^{(i)}, \dots, r_{t,m}^{(i)})$ the t-th sample in $\bar{D}_i$. Further define

$$\operatorname{bias}_A(i) = \frac{1}{n_i}\sum_{t=1}^{n_i}\left[\max_m v(\pi_m, x_t^{(i)}) - \mathop{\mathbb{E}}_{r_1,\dots,r_m|x=x_t^{(i)}}[\max_m r_m]\right]$$

and

$$Y_{i,t} = \mathop{\mathbb{E}}_{r_1,\dots,r_m|x=x_t^{(i)}}[\max_m r_m] - \max_m r_{t,m}^{(i)} - v(\pi_f, x_t^{(i)}) + \sum_{m=1}^{k} r_{t,m}^{(i)} f(m|x_t^{(i)})$$

Then we can decompose the difference in losses as

$$\mathcal{L}_A(f,\lambda) - \widehat{\mathcal{L}}_A(f,\lambda) = \sum_{i=1}^{k} \lambda_i \, \mathrm{bias}_A(i) + \sum_{i=1}^{k} \frac{\lambda_i}{n_i} \sum_{t=1}^{n_i} Y_{i,t}.$$

Since $Y_{i,t}$ are all independent from each other, we can bound the second term using concentration arguments as

$$\sum_{i=1}^{k} \frac{\lambda_i}{n_i} \sum_{t=1}^{n_i} Y_{i,t} \le O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log(1/\delta)}{n_i}} + \max_i \frac{\lambda_i \log(1/\delta)}{n_i} \right)$$

with probability at least $1 - O(\delta)$. Note that we can bound the negative, $-\sum_{t=1}^{n_i} Y_{i,t}$ analogously. Further, by Jensen's inequality, $\mathrm{bias}_A(i) \le 0$ for all i. Combining these bounds yields the desired statement. □

Theorem 3 (Regret bound for Option A)**.** *Assume that the function class $\mathcal{F}$ is convex. Then the solution $\bar{f}$ returned by Algorithm 1 with Option A satisfies with probability at least $1 - O(\delta)$*

$$\mathrm{reg}(\pi_{\bar{f}}, D_\lambda) \le \mathrm{reg}(\pi_{pt}, D_\lambda) + \widehat{V}_A^\star + \frac{\mathrm{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right) \tag{15}$$

$$+ O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log(|\mathcal{F}|/\delta)}{n_i}} + \max_i \frac{\lambda_i \log(|\mathcal{F}|/\delta)}{n_i} \right) \tag{16}$$

Further, if there exists an $f \in \mathcal{F}$ which perfectly predicts the maximum score per sample, i.e., $\sum_{i=1}^{k} \mathbb{E}_{(x,r_1,\dots,r_k)\sim\bar{D}_i}[\max_m r_m] = \sum_{i=1}^{k} \mathbb{E}_{(x,r_1,\dots,r_k)\sim\bar{D}_i} \mathbb{E}_{j\sim f(x)} r_j$, then $\widehat{V}_A^\star \le 0$. If this only holds on a population level and for expected scores, i.e., $\sum_{i=1}^{k} \mathbb{E}_{x\sim D_i} \max_m v(\pi_m, x) = \sum_{i=1}^{k} \mathbb{E}_{x\sim D_i} v(\pi_f, x)$, then we can still bound $\widehat{V}_A^\star \le \max_i |\mathrm{bias}_A(i)| + O\left(\frac{\log(|\mathcal{F}|/\delta)}{\sqrt{\min_i n_i}}\right)$.

Proof. We can decompose the regret of $\bar{f}$ on D_λ as

$$\begin{aligned}
\mathrm{reg}(\pi_{\bar{f}}, D_\lambda) &= \max_{\pi\in\Pi} v(\pi, D_\lambda) - \mathop{\mathbb{E}}_{x\sim D_\lambda}\left[\max_m v(\pi_m, x)\right] + \mathop{\mathbb{E}}_{x\sim D_\lambda}\left[\max_m v(\pi_m, x)\right] - v(\pi_{\bar{f}}, D_\lambda) \\
&= \max_{\pi\in\Pi} v(\pi, D_\lambda) - \mathop{\mathbb{E}}_{x\sim D_\lambda}\left[\max_m v(\pi_m, x)\right] + \mathcal{L}_A(\bar{f}, \lambda) \\
&\le \max_{\pi\in\Pi} v(\pi, D_\lambda) - \mathop{\mathbb{E}}_{x\sim D_\lambda}\left[\max_m v(\pi_m, x)\right] + \widehat{\mathcal{L}}_A(\bar{f}, \lambda) \\
&\quad + O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log(|\mathcal{F}|/\delta)}{n_i}} + \max_i \frac{\lambda_i \log(|\mathcal{F}|/\delta)}{n_i} \right) && \text{(Lemma 7)}
\end{aligned}$$

To obtain a bound on $\widehat{\mathcal{L}}_A(\bar{f}, \lambda)$, we apply the game-theoretic arguments from Lemma 4

$$\widehat{\mathcal{L}}_A(\bar{f}, \lambda) \le \widehat{V}_A^\star + \frac{\mathrm{Reg}_{\mathcal{F}}(T)}{T} + O\left(\sqrt{\frac{\log(k|\mathcal{F}|/\delta)}{T}}\right)$$

and it only remains to control the optimal value of the game $\widehat{V}_A^\star$.

$$\begin{aligned}
\widehat{V}_A^\star &= \max_{\lambda\in\Delta_k} \inf_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(x,r_1,\dots,r_k)\sim\bar{D}_\lambda} \mathop{\mathbb{E}}_{j\sim f(x)}[\max_m r_m - r_j] \\
&\le V_A^\star + \max_{\lambda\in\Delta_k}\left\{ \left|\sum_{i=1}^{k} \lambda_i \, \mathrm{bias}_A(i)\right| + O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log(|\mathcal{F}|/\delta)}{n_i}} + \max_i \frac{\lambda_i \log(|\mathcal{F}|/\delta)}{n_i} \right) \right\} \\
&\le V_A^\star + \max_i |\mathrm{bias}_A(i)| + O\left(\frac{\log(|\mathcal{F}|/\delta)}{\sqrt{\min_i n_i}} \right)
\end{aligned}$$

□

B.3 Alternate Oracles

In this section we consider replacing the linear losses, $\ell_t^{(i)}$, from Algorithm 1 with a log-loss. Such a choice is natural whenever we consider $\mathcal{F}$ to be some family of Transformer networks for which modern ML packages use optimizers tailored to the cross-entropy loss. The losses constructed by Algorithm 1 are log-losses and so we need a different version of the Online Learning Oracle which we defined below.

Definition 2 (Online learning logistic oracle)**.** *An algorithm* OLLO *is called a online learning oracle for a class* $\mathcal{F} \subseteq \mathcal{X} \to \Delta_k$ *if the following holds. Let* $(x_1, \ell_1, \ldots, x_T, \ell_T)$ *be an arbitrary, possibly adversarial sequence of contexts and loss pairs.* OLLO *observes* x_t *sequentially and maintains a sequence of policies* f_t *which it updates by observing the loss vector* ℓ_t*. The total regret of* OLLO

$$\mathrm{Reg}_{\mathcal{F}}^{\mathsf{OLLO}}(T) = \max_{f \in \mathcal{F}} \sum_{t=1}^{T} \langle \log f(x_t) - \log f_t(x_t), \ell_t \rangle = o(T).$$

is sublinear with high probability, at least $1 - \delta$.

The problem of Online Logistic Regression has been extensively studied in the online learning literature (Kakade and Ng, 2004; Xiao, 2009; McMahan and Streeter, 2012; Hazan et al., 2014; Foster et al., 2018; Shamir, 2020). Using OLLO we can instantiate a new version of Algorithm 1 with the following losses for the min-player $\ell_t^{(i)\prime} = -\lambda_{t,i} e_{y_t^{(i)}}$ where $y_t^{(i)} \in \{j \in [k] : r_{t,j}^{(i)} = c_t^{(i)}\}$. Option A and Option B then correspond to the following two choices of $y_t^{(i)}$

$$y_t^{(i)} = \begin{cases} \mathrm{argmax}_{j \in [k]} \, r_{t,j}^{(i)} & \text{Option A} \\ r_{t,i}^{(i)} & \text{Option B.} \end{cases}$$

Next, we prove the counterpart to Lemma 4 for the classifier setting.

Lemma 8. *For any* $\lambda \in \Delta(k)$ *it holds that*

$$\begin{aligned} &\sum_{t=1}^{T} \sum_{i=1}^{k} \lambda_i \left(c_t^{(i)} - \sum_{j=1}^{k} f_t(j|x_t^{(i)}) r_{t,j}^{(i)} \right) \\ &\leq \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \sum_{i=1}^{k} -\lambda_{t,i} \log(f(y_{t,i}^{(i)}|x_t^{(i)})) + \mathrm{Reg}_{\mathcal{F}}^{\mathsf{OLLO}}(T) + O(\sqrt{kT \log(k|\mathcal{F}|/\delta)}), \end{aligned}$$

with probability $1 - O(\delta)$.

Proof. The definition of OLLO together with the standard analysis for the regret of the max-player imply the following holds with probability $1 - O(\delta)$

$$\begin{aligned} &\sum_{t=1}^{T} \sum_{i=1}^{k} -\lambda_{t,i} \left(\log(f_t(y_t^{(i)}|x_t^{(i)})) - \log(f(y_t^{(i)}|x_t^{(i)})) \right) \leq \mathrm{Reg}_{\mathcal{F}}^{\text{classifier}}(T) \\ &\sum_{t=1}^{T} \sum_{i=1}^{k} \lambda_i \left(c_t^{(i)} - \sum_{j=1}^{k} f_t(j|x_{i,t}) r_{t,j}^{(i)} \right) \\ &- \sum_{t=1}^{T} \sum_{i=1}^{k} \lambda_{t,i} \left(c_t^{(i)} - \sum_{j=1}^{k} f_t(j|x_{i,t}) r_{t,j}^{(i)} \right) \leq O(\sqrt{kT \log(k|\mathcal{F}|/\delta)}) \end{aligned}$$

And so for any fixed $\lambda \in \Delta(k)$ we have

$$\begin{aligned} &\sum_{t=1}^{T} \sum_{i=1}^{k} \lambda_i \left(c_t^{(i)} - \sum_{j=1}^{k} f_t(j|x_t^{(i)}) r_{t,j}^{(i)} \right) \\ &\leq \sum_{t=1}^{T} \sum_{i=1}^{k} \lambda_{t,i} \left(c_t^{(i)} - \sum_{j=1}^{k} f_t(j|x_t^{(i)}) r_{t,j}^{(i)} \right) + O(\sqrt{kT \log(k|\mathcal{F}|/\delta)}) \\ &\leq \sum_{t=1}^{T} \sum_{i=1}^{k} \lambda_{t,i} r_{t,y_t^{(i)}}^{(i)} \left(1 - f_t(y_t^{(i)}|x_t^{(i)}) \right) + O(\sqrt{kT \log(k|\mathcal{F}|/\delta)}) \end{aligned}$$

$$\leq \sum_{t=1}^{T}\sum_{i=1}^{k} -\lambda_{t,i} r^{(i)}_{t,y_t^{(i)}} \log(f_t(y_t^{(i)}|x_t^{(i)})) + O(\sqrt{kT\log(k|\mathcal{F}|/\delta)}).$$

for any i, where the last inequality uses $1-x \leq -\log(x), x \in [0,1]$. The min-player regret guarantee together with the fact that $r^{(i)}_{t,y_t^{(i)}} \in [0,1]$ imply

$$\sum_{t=1}^{T}\sum_{i=1}^{k} \lambda_i \left(c_t^{(i)} - \sum_{j=1}^{k} f_t(j|x_t^{(i)}) r^{(i)}_{t,j}\right)$$
$$\leq \min_{f\in\mathcal{F}} \sum_{t=1}^{T}\sum_{i=1}^{k} -\lambda_{t,i}\log(f(y_{t,i}^{(i)}|x_t^{(i)})) + \mathrm{Reg}_{\mathcal{F}}^{\mathsf{OLLO}}(T) + O(\sqrt{kT\log(k|\mathcal{F}|/\delta)}).$$

□

We need the following assumption to guarantee boundedness of the log-loss for the concentration argument.

Assumption 1. $\forall f \in \mathcal{F}$ and for any $(y,x) \in \mathcal{Y}\times\mathcal{X}$ it holds that $f(y|x) \geq \frac{1}{T}$.

Lemma 9. *Under Assumption 1 it holds that*

$$\mathbb{P}\left(\sum_{t=1}^{T}\sum_{i=1}^{k} -\lambda_{t,i}\left(\log(f(y_t^{(i)}|x_t^{(i)})) - \mathbb{E}[\log(f(y_t^{(i)}|x_t^{(i)}))]\right) \geq \sqrt{2\log(T)\log(|\mathcal{F}|/\delta)}\right) < \delta.$$

Proof. Directly follows from Azuma-Hoeffding and the boundedness of the log-loss under the assumption. □

In Appendix D we present a concentration bound for unbounded losses with bounded second moment which can be applied instead of Lemma 9. Combining the two lemmas gives us the following result.

Theorem 4. *Under Assumption 1 with probability* $1-\delta$ *it holds that for any* $\lambda \in \Delta(k)$ *and* $f \in \mathcal{F}$

$$\mathbb{E}_{i\sim\lambda, x\sim D_i}[c^{(i)} - \langle \bar{f}, r^{(i)}\rangle] \leq \mathbb{E}_{i\sim\bar{\lambda}, x^{(i)}\sim\bar{D}_i}[-\log(f(y^{(i)}|x^{(i)}))]$$
$$+ \frac{\mathrm{Reg}_{\mathcal{F}}^{\mathsf{OLLO}}(T)}{T} + O\left(\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{T}} + \sqrt{\frac{\log(k/\delta)}{T}}\right)$$
$$+ O\left(\sqrt{\sum_{i=1}^{k}\frac{\lambda_i^2\log(k|\mathcal{F}|T/\delta)}{n_i}} + \max_i \frac{\lambda_i\log(k|\mathcal{F}|T/\delta)}{n_i}\right)$$

Proof. We begin by arguing that

$$\mathbb{E}_{i\sim\lambda, x\sim\bar{D}_i}[c^{(i)} - \langle \bar{f}, r^{(i)}\rangle] \leq \mathbb{E}_{i\sim\bar{\lambda}, x^{(i)}\sim\bar{D}_i}[-\log(f(y^{(i)}|x^{(i)}))]$$
$$+ \frac{\mathrm{Reg}_{\mathcal{F}}^{\mathsf{OLLO}}(T)}{T} + O\left(\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{T}} + \sqrt{\frac{\log(k/\delta)}{T}}\right)$$

This holds as follows. We combine the regret bound from Lemma 8 together with the concentration of Lemma 2 and Lemma 9.

Finally, we convert the LHS of the above lemma to a concentration over the population $\mathbb{E}_{i\sim\lambda,x\sim D_i}[c^{(i)} - \langle \bar{f}, r^{(i)}\rangle]$ as follows. First note that for any fixed $f \in \mathcal{F}$:

$$\mathbb{E}_{i\sim\lambda,x\sim D_i}[c^{(i)} - \langle f, r^{(i)}\rangle] = \sum_{i=1}^{k}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i} c_j^{(i)} - \langle f, r_j^{(i)}\rangle.$$

We can then argue as in Lemma 2 that for all $\lambda \in \Delta(k)$ uniformly it holds that

$$\sum_{i=1}^{k}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\mathbb{E}[c_j^{(i)} - \langle f, r_j^{(i)}\rangle] - \sum_{i=1}^{k}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i} c_j^{(i)} - \langle f, r_j^{(i)}\rangle \leq O\left(\sqrt{\sum_{i=1}^{k}\frac{\lambda_i^2\log(k/\delta)}{n_i}} + \max_i\frac{\lambda_i\log(k/\delta)}{n_i}\right),$$

w.p. $1 - O(\delta)$, where we use Bernstein's inequality instead of Hoeffding's inequality. An additional union bound over $\mathcal{F}$ now implies

$$\mathbb{P}\Bigg(\sup_{\lambda\in\Delta(k), f\in\mathcal{F}} \sum_{i=1}^{k} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \mathbb{E}[c_j^{(i)} - \langle f, r_j^{(i)}\rangle] - \sum_{i=1}^{k} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} c_j^{(i)} - \langle f, r_j^{(i)}\rangle$$

$$\geq \Omega\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log(k|\mathcal{F}|/\delta)}{n_i}} + \max_i \frac{\lambda_i \log(k|\mathcal{F}|/\delta)}{n_i}\right)\Bigg) \leq \delta.$$

Finally, we note that $\bar{f} \in \mathcal{F}$ by convexity of $\mathcal{F}$. and thus we need an extra union bound over T. This completes the proof of the theorem. □

We can now show counterparts to Theorem 3 and Theorem 2.

Corollary 2. *For any convex $\mathcal{F}$ for which Assumption 1 holds we have that for all $\lambda \in \Delta(k)$ with probability $1 - \delta$*

$$\mathrm{reg}(\pi_{\bar{f}}, D_\lambda) \leq \min_{f\in\mathcal{F}} \mathbb{E}_{i\sim\bar{\lambda}, x^{(i)}\sim\bar{D}_i}[-\log(f(y^{(i)}|x^{(i)}))]$$

$$+ \frac{\mathrm{Reg}_{\mathcal{F}}^{\mathit{OLLO}}(T)}{T} + O\left(\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{T}} + \sqrt{\frac{\log(k/\delta)}{T}}\right)$$

$$+ O\left(\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2 \log(k|\mathcal{F}|T/\delta)}{n_i}} + \max_i \frac{\lambda_i \log(k|\mathcal{F}|T/\delta)}{n_i}\right),$$

where for Option A we have $y^{(i)} = \mathrm{argmax}_{y\in[k]} r^\star(y, x^{(i)})$ and for Option B we have $y^{(i)} = i$.

Proof. The definition of regret for Option A implies that

$$\begin{aligned}
\mathrm{reg}(\pi_{\bar{f}}, D_\lambda) &= v(\pi_A^\star, D_\lambda) - v(f, D_\lambda) = \mathbb{E}_{i\sim\lambda, x\sim D_i}[\max_{j\in[k]} v(\pi_j, x^{(i)}) - v(\bar{f}, x^{(i)})] \\
&\leq \mathbb{E}_{i\sim\lambda, x\sim D_i}[\operatorname*{argmax}_{y\in[k]} r^\star(y, x^{(i)}) - \langle \bar{f}, r^\star(\cdot, x^{(i)})\rangle] \\
&= \mathbb{E}_{i\sim\lambda, x\sim D_i}[c^{(i)} - \langle \bar{f}, r^{(i)}\rangle].
\end{aligned}$$

The bound now follows from Theorem 4. For Option B we have a similar derivation with

$$\begin{aligned}
\mathrm{reg}(\pi_{\bar{f}}, D_\lambda) &= v(\pi_A^\star, D_\lambda) - v(\pi_{\bar{f}}, D_\lambda) = \mathbb{E}_{i\sim\lambda, x\sim D_i}[v(\pi_i, x^{(i)}) - v(\pi_{\bar{f}}, x^{(i)})] \\
&= \mathbb{E}_{i\sim\lambda, x\sim D_i, j\sim\pi_i(x^{(i)})}\left[r^\star(j, x^{(i)}) - \sum_{l=1}^{k}\sum_{s=1}^{k} \bar{f}(s|x^{(i)})\pi_s(l|x^{(i)}) r^\star(l, x^{(i)})\right] \\
&= \mathbb{E}_{i\sim\lambda, x\sim D_i, j\sim\pi_i(x^{(i)})}[c^{(i)} - \langle \bar{f}, r^{(i)}\rangle].
\end{aligned}$$

The bound again follows from Theorem 4. □

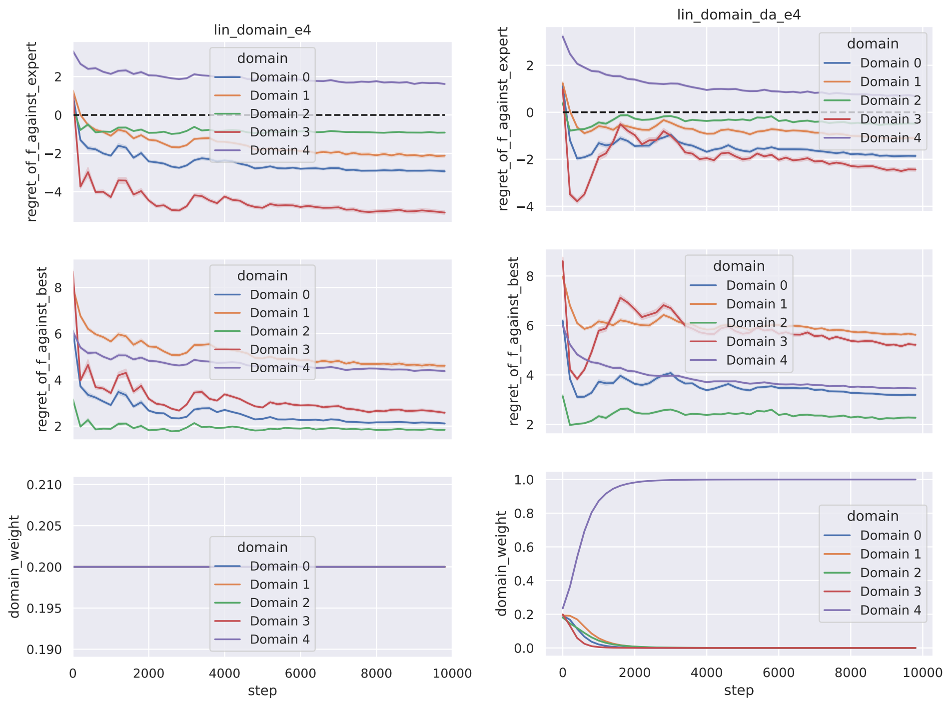

Figure 1: Comparison of Algorithm 1 with Option B (right) against non-robust version (left).

C Experimental results

In Figure 1, we present the results for Algorithm 1 under Option A. The online game runs for 1000 iterations, with each iteration mini-batched using a size of 256. The first row of the figure illustrates the regret of our algorithm compared to a competitor that always selects the best expert for each domain. The regret for domain D_i is defined as the difference between the reward of $\pi_i(x^{(i)})$ and the reward obtained by our domain classifier, which integrates all π_i for $i \in [k]$. Notably, the domain adaptation approach (on the right) consistently outperforms or maintains performance equivalent to the best domain expert model across all five domains.

In the second row, we show the regret against the pointwise best policy for each input x, represented as the regret relative to $\max_{j \in k} r_{t,j}^{(i)}$. Again, our method surpasses the simple domain classifier. Lastly, the third row displays the domain weights returned by the two approaches: while the domain classifier does not update the domain weights, the max-player in Algorithm 1 significantly increases the weight of domain D_4.

D Unbounded loss bound

The following generalization bound follows directly Theorem 3 of (Cortes et al., 2021a). It holds for any unbounded loss function with bounded second-moment. In particular, it can be applied to the log loss when the second-moment is bounded.

Theorem 5. *Fix $\varepsilon \in (0, 1)$. Then, for any hypothesis set $\mathcal{H}$ such that $\mathbb{E}_{x \sim \mathcal{D}}[\ell^2(h, x)] < +\infty$ for all $h \in \mathcal{H}$, the following holds with probably at least $1 - \delta$ over the draw of a sample of size m from $\mathcal{D}$:*

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}}[\ell(h, x)] - \mathop{\mathbb{E}}_{x \sim S}[\ell(h, x)] \leq \gamma \sqrt{\mathop{\mathbb{E}}_{x \sim \mathcal{D}}[\ell^2(h, x)] \frac{\Delta_m}{m}} + \varepsilon,$$

where $\Delta_m = \log \mathbb{E}[\mathcal{N}_\infty(\ell(\mathcal{H}), \frac{\varepsilon}{2}, x_1^{2m})] + \log \frac{1}{\delta}$, $\gamma = \Gamma_0\left(\sqrt{\frac{\Delta_m}{m}}\right) = \mathcal{O}(\log m)$, and $\Gamma_0(\mu) = \frac{1}{2} + \sqrt{1 + \frac{1}{2} \log \frac{1}{\mu}}$ for any $\mu > 0$. $\mathcal{N}_\infty(\ell(\mathcal{H}), \frac{\varepsilon}{2}, x_1^{2m})$ represents the ℓ_∞-covering number of the ℓ-losses

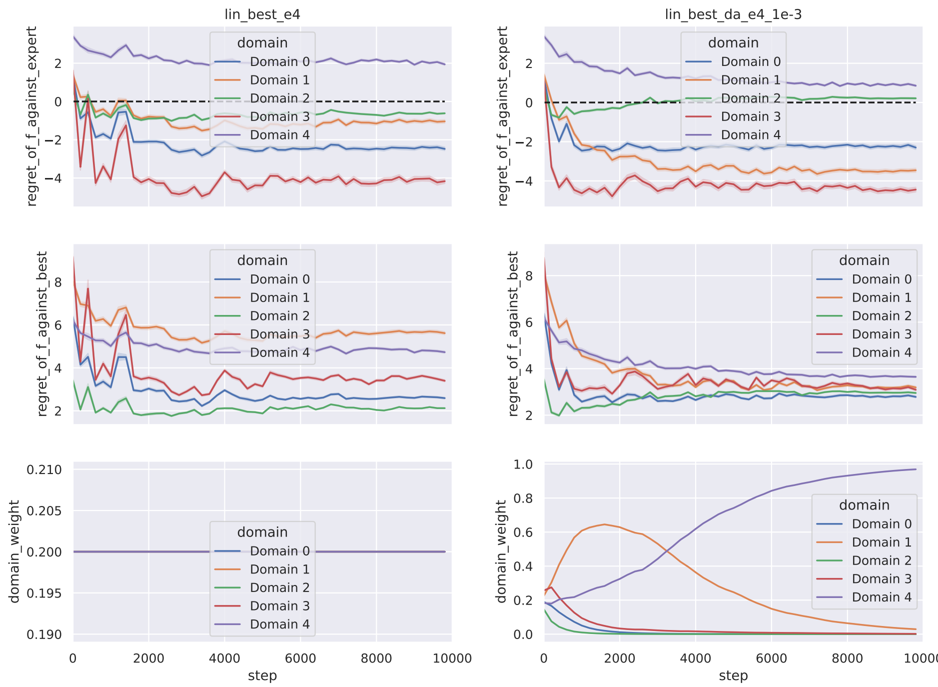

Figure 2: Comparison of Algorithm 1 with Option A (right) against non-robust version (left).

associated with the hypotheses in $\mathcal{H}$ based on a sample of size $2m$, denote by x_1^{2m}, with a precision of $\frac{\varepsilon}{2}$.

In particular, we can choose $\varepsilon = \frac{1}{m}$ in the bound. The result generalizes to the case where only a higher-order moment of the loss (higher than 2) is bounded.