A THEORETICAL STUDY OF MODEL ROUTING WITH DISTRIBUTIONAL LOSSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative AI systems increasingly rely on routing user prompts among multiple generative models with different capabilities, costs, and output qualities. This is the problem of model routing, where a decision-maker selects among different generative models in an online fashion to optimize the quality and minimize the cost of the generated corpus. Currently, the majority of model routing algorithms use hard-coded routing choices or assume that the generated output can be evaluated in a per-prompt manner. However, a per-prompt loss alone cannot capture the efficacy of a generative model; instead, one needs to evaluate the generated output at a *distributional* level. For example, the Fréchet Inception Distance (FID) depends on the distribution of generated images.

Motivated by this example, we introduce a formulation of the online model routing problem where the loss can depend on both per-round and distributional-level quantities. We show that one can extend classical decision-making algorithms such as Upper Confidence Bound (UCB) and Explore-Then-Exploit to handle such losses when the loss is convex and satisfies a tail-bound assumption. These conditions are satisfied by canonical examples used widely in practice such as FID and Frontier Integral (FI). Overall, our work develops a principled formulation of model routing with novel distributional losses and shows that principled decision making is possible in these challenging settings.

1 INTRODUCTION

We consider the online model routing problem: given a user's prompt and several generative models, how can we optimally route the prompt by routing the user's prompt to the best possible model? Here, the notion of 'best' in the choice of model can capture several goals: the model that generates the closest output to the user's prompt, the model that generates an output so the distribution of all generated outputs faithfully represents a given distribution, or the model that produces an output with minimal cost.

The model routing problem is an increasingly important problem with major economic impact; see e.g. (Jitkrittum et al., 2025; Moslem & Kelleher, 2026; Chadli et al., 2025) for an incomplete list. A clear motivation is the prohibitively high inference cost of LLMs (Chadli et al., 2025; Wan et al., 2024; Zhou et al., 2024). One may hope that one can learn to occasionally use high-cost LLMs with higher fidelity output, while more frequently using low-cost LLMs whose generation to the user's prompt is sufficient. Beyond cost, there is growing evidence that the output of different generative models have different strengths on different types of generation tasks, see e.g. (Bommasani et al., 2023; Srivastava et al., 2023; Jiang et al., 2024; Chiang et al., 2024; Zheng et al., 2023). Thus given a prompt, one aims to choose a generative model whose strengths best match the prompt.

In addition to the aforementioned considerations, the notion of 'best' model needs to take into account several other factors. Perhaps most notably, the notion of 'best' model often depends on the corpus of generated outputs rather than the quality or cost of a single generation for a given prompt. Consider the example of image generation. It is possible that a single image may be a effective enough response to many prompts. As a simplistic example, an image of a red car on Golden Gate Bridge would be a perfectly valid response to the prompts: 'Generate a picture of a red car', 'Generate a picture of a car on a bridge', 'Generate a picture of a red car in the Bay Area'. Yet repeatedly producing a single image of a red car on Golden Gate Bridge is certainly not an effective response over the set of all

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

prompts, nor does it resemble realistic output of generative models. A more natural goal would be at the *distributional* level: for the distribution of generated images over all the prompts to resemble the set of natural images or a large corpus of images such as ImageNet (Deng et al., 2009).

Goals and High Level Formulation. In this paper, we introduce a formulation of the online model routing problem where the loss can depend on both per-round and distributional-level quantities. Taking inspiration from the bandit literature (see e.g. Bubeck & Cesa-Bianchi (2014); Cesa-Bianchi & Lugosi (2006); Lattimore & Szepesvári (2020)), we consider K generative models $m_1, \ldots, m_K$, corresponding to 'arms' in the bandit literature. Each model m_i takes in a prompt $p \in \mathcal{P}$ and produces a generated output $x \in \mathcal{X}$ (possibly randomized) where $\mathcal{P}$ denotes the space of prompts and $\mathcal{X}$ denotes the space of outputs. Thus $m_i : \mathcal{P} \to \Delta(\mathcal{X})$ where $\Delta(\mathcal{X})$ denotes the space of measures on $\mathcal{X}$.

The learner aims to choose the models $m_1, \ldots, m_K$ in an online fashion, such that their performance is similar to that of the best model in hindsight, similar to the notation of *regret* from the bandit literature. This captures the notion of routing prompts to maximize the quality and minimize the cost of the generated corpus, which is measured here relative to a comparator. The comparator in the definition of regret varies per application and we provide a precise definition and its motivation for each application in Section 3, Section 4.

The learning procedure now proceeds as follows for rounds $1 \le t \le T$, where T denotes the total number of rounds of generation:[1]

1. A prompt p_t arrives from the users (henceforth, 'user').
2. Learner picks a generative model m_{i_t} indexed by $i_t \in [K]$.
3. Learner provides the user with the generation $x_t = m_{i_t}(p_t)$.

For all $1 \le t \le T$, we define the empirical measure

$$\{(p_1, x_1), \ldots, (p_t, x_t)\} := \hat{\mu}_T. \tag{1}$$

At the end of the T rounds, we wish to establish measure the quality of the corpus of generated outputs $\hat{\mu}_T$ outputted to the user. In practice, the quality of $\hat{\mu}_T$ is measured by a range of losses, which broadly fall under three categories:

1. Losses that measure mismatch between the generated output to the prompt. This covers examples such as CLIPScore for image-to-text (Hessel et al., 2021), HPSv2 for text-to-image (Wu et al., 2023), VQAScore for text-to-image/video/3D (Lin et al., 2024), pass@k for code (Chen et al., 2021), and many more.
2. Losses that measure the statistical distance of $\hat{\mu}_T$ with respect to a given reference measure. This covers prototypical distributional losses used ubiquitously in practice such as Fréchet Inception Distance (FID) for images (Heusel et al., 2017a), MAUVE/FI in text (Pillutla et al., 2021; Liu et al., 2021), Fréchet BERT Distance for text (Alihosseini et al., 2019), Precision–Recall for generative models for images and general domains (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), MMD in embedding space for general domains (Gretton et al., 2012), and many more.
3. Losses that measure the diversity of $\hat{\mu}_T$. This covers examples such as the Vendi Score for general domains (Friedman & Dieng, 2023), Distinct-n and Self-BLEU for text (Li et al., 2016; Zhu et al., 2018), Inception Score for images (Salimans et al., 2016), Behavioral diversity for code (Chen et al., 2021), and many more.

We emphasize that losses of category 2) and 3) are necessary to ensure that the generated outputs resemble a realistic distribution. Again consider the example of an image of a red car on Golden Gate Bridge; repeatedly generating this image certainly is a valid response to many possible prompts, but this not reasonable behavior for a generative model. This highlights the necessity of losses such as FID in categories 2), 3) above, which are *global* quantities that depends on the entire empirical distribution $\hat{\mu}_T$. This is in contrast to the round-by-round loss that is typically studied in the bandit literature (Bubeck & Cesa-Bianchi, 2014; Cesa-Bianchi & Lugosi, 2006; Lattimore & Szepesvári,

[1] In this paper, we assume the number of rounds is known in advance for simplicity; more on this is in Section 2.

2020). Such global losses pose significant challenges to the typical style of algorithms used in the classical bandit setting (for instance inverse propensity scores).

In our formulation, we generalize all of these examples as follows. After all T rounds of generation, the learner suffers a *total* loss

$$L_{\text{global}}(\hat{\mu}_T) + \frac{1}{T}\sum_{t=1}^{T} \ell_{\text{round}}(p_t, x_t) = L_{\text{global}}(\hat{\mu}_T) + \mathbb{E}_{(p_t,x_t)\sim\hat{\mu}_T}[\ell_{\text{round}}(p_t, x_t)] \\ := L_{\text{total}}(\hat{\mu}_T), \tag{2}$$

where $L_{\text{global}}(\hat{\mu}_T)$ is the *global* loss depending on the corpus of generated outputs $\hat{\mu}_T$ and each $\ell_{\text{round}}(p_t, x_t)$ is the *local* loss per round. We note the local loss can capture losses in category 1) above that measure the mismatch between the generated output and the prompt, but also can capture economic considerations such as the cost of generating output x_t with model m_{i_t}. Note the relative weight between the local and global loss can be absorbed into L_{global} and ℓ_{round}.

Our goal is then to minimize the *regret*: the difference between the total loss (2) and the total loss of a suitable comparator.

Remark 1. *Observe a loss of the form (2) subsumes the loss from the bandit literature where $L_{\text{global}} \equiv 0$. Thus all lower bounds from the bandit literature carry over to our formulation.*

Remark 2 (On choosing a generative model more than once per round)**.** *We emphasize that in our formulation, for each round t, we can choose another generative model $m_{i'_t}$ and observe the generation $x'_t = m_{i'_t}(p_t)$ without suffering the associated loss if the cost of generation is minimal in ℓ_{round}. In particular, the associated loss is not suffered if it is not the generation returned to the user. This x'_t can be used to provide additional information by the learning algorithm.*

We believe the assumption in Remark 2 is sensible in situations where the models are low cost. If we do not account for cost, the associated loss with a generation is not suffered if it is not the generation returned to the user. This is in contrast with the bandit literature (Bubeck & Cesa-Bianchi, 2014; Cesa-Bianchi & Lugosi, 2006; Lattimore & Szepesvári, 2020): for example, in online advertising one must suffer the loss from displaying an ad, and in a slot machine one must commit to the arm pulled.

Summary of Contributions. Our contributions are now as follows:

1. Section 3: Consider when the prompts p_t are drawn i.i.d. and L_{global} satisfies Assumption 2 (concentration of L_{global}), Assumption 3 (convexity of L_{global}), Assumption 4 (boundedness of L_{global}). We establish that a variant of the classic UCB algorithm (Auer et al., 2002a) achieves vanishing regret, where regret is measured against the best model in hindsight. In particular, when Assumption 2 holds with $p = 2$, we establish a $\tilde{O}(T^{-1/2})$ regret bound.[2] Note that this allows for interactive workflows. Specifically, if we view the user intent as i.i.d., then the model outputs can implicitly account for multiple user interactions and our results still hold.
2. Section 4: Consider when we have a set of lightweight, low cost models with different strengths on different generation tasks, and a class $\mathcal{F}$ of routing functions mapping prompts to models (e.g., functions in $\mathcal{F}$ can map prompts to the model that best suits the associated generation task). Suppose that the assumptions from Section 3 also hold. We establish that a variant of the Explore-then-Exploit algorithm (Lattimore & Szepesvári, 2020) achieves vanishing regret, where regret is measured against the model chosen by the best routing function in hindsight (which may vary across the different prompts).

   Furthermore, consider when Assumption 2 holds with $p = 2$, and also when ℓ_{round} does not take into account cost of generation as per Remark 2 – we believe this is a reasonable assumption here as this setting is motivated by situations where the models are lightweight and low cost. In this case, we establish a $\tilde{O}(T^{-1/2})$ regret bound. We also provide supporting experiments in Section 5.
3. Section C: We prove that when the prompts p_t are drawn adversarially, effective decision making – in the sense of achieving vanishing regret – is not possible. This justifies the

[2] In this paper, $\tilde{O}$ hides logarithmic factors; explicit dependence on log factors are given in each of our Theorems.

choice of i.i.d. prompts p_t in Section 3, Section 4. We also give assumption-specific lower bounds for Assumption 2 and Assumption 3.

The notions of regret defined above are motivated in Section 3, Section 4. Here Assumption 2 asserts that high-probability concentration of the loss L_{total} with independent samples holds, and Assumption 3 asserts that L_{total} is convex in its argument (which is a measure). These assumptions and their motivations are discussed further in Section 2.

Related work: As discussed before, model routing is an important problem of major impact. As such, there has been a large body of recent empirical work on this problem, which is too large in scope to completely cover here. See (Chadli et al., 2025) for a representative survey. Some of these works on model routing involve bandit-based methods, for an incomplete list see (Nguyen et al., 2024; Wang et al., 2025; Panda et al., 2025; Ziller et al., 2026; Chiang et al., 2025; Xia et al., 2024; Luo et al., 2024; Qin et al., 2024; Frick et al., 2025). More directly related to our work, there recently have been several works studying model routing via bandit-style approaches similar to our work and proving associated regret bounds. Specifically, see the works Hu et al. (2024; 2025a;b); Rezaei et al. (2024). We also mention that there are related works on online learning with global losses, discussed in more detail in Section 2.

We emphasize that Hu et al. (2025a;b); Rezaei et al. (2024) study the model routing problem, but *without a prompt*. In contrast, in our setting, we receive prompts from users and then aim to route to generative models. We also note that the regret bound in Hu et al. (2025b) obtains a regret bound of the form $O(T^{-1/4})$ for the FID loss. Specifically, see eq. (3) therein, which contains a nested square root of the form $\sqrt{\sigma_g^2\|\Sigma_g\|\sqrt{\frac{r(\Sigma_g)}{T}}+\frac{\mathcal{I}_g}{T}}$ (note Hu et al. (2025b) use n rather than T). By contrast, our Theorem 1 and Theorem 3 obtain a regret bound of the form $\tilde{O}(T^{-1/2})$ for the FID loss. This is a consequence of our concentration result for FID discussed in Example 1.

The problem of routing to generative models with a prompt using bandit style algorithms is studied in Hu et al. (2024). However, their loss is a local, per-round quantity; specifically, this is their score s_t in round t of the algorithm, see Algorithm 2 therein. Instead, we do not receive a score per round, and L_{total} that we study here is a *global* quantity that can only be computed at the end of the generation protocol. This is in line with many scoring metrics of generative models such as FID, MAUVE/FI, and many others, which as discussed earlier are global losses.

2 The Learning Setting

Here we complete the formalization of our learning setting. Throughout this paper, we will assume that the prompts p_t are drawn i.i.d.:

Assumption 1. *The prompts p_t are drawn i.i.d. from a distribution $\mathcal{D}_{\text{prompts}}$ over the space $\mathcal{P}$.*

The assumption of i.i.d. prompts may seem strong, however we demonstrate in Section C that non-trivial results are impossible with adversarially generated prompts. Furthermore, the abstraction of i.i.d. prompts can model not just a single prompt, but an entire interaction between a user and a fixed model. Each interaction begins with a i.i.d. user intent, and any follow-up prompts are treated as part of that same interaction rather than as separate draws.

Since each $m_i : \mathcal{P} \to \Delta(\mathcal{X})$ and the prompts are i.i.d., it follows that each model m_i induces the following joint distribution μ_i over $\mathcal{P} \times \mathcal{X}$ defined as follows:

$$\mu_i := (p, m_i(p)) \text{ where } p \sim \mathcal{D}_{\text{prompts}} \quad \text{for each} \quad i \in [K]. \tag{3}$$

We consider T rounds of generation where the prompts $p_t \sim \mathcal{D}_{\text{prompts}}$, as discussed in Section 1. For simplicity, we assume that T is known in advance; however, one can relax this assumption using the doubling trick, see Section 6.2 of Lattimore & Szepesvári (2020).

To handle the distribution-level loss L_{total} that we consider in this paper, we will impose the following assumptions. As discussed next, we believe Assumption 2 is very natural, while Assumption 3 is satisfied be a range of losses used in practice. In fact, we show their necessity in Section C. We explicitly verify these assumptions for the examples of FID and FI in Example 1, Example 2.

Assumption 2 (Concentration of L_{total})**.** *Let μ be a given probability distribution on $\mathcal{P} \times \mathcal{X}$. Suppose $(p_1, x_1), \ldots, (p_n, x_n) \sim \mu$ are drawn i.i.d., and let $\hat{\mu}_n$ denote the empirical measure on $\{(p_1, x_1), \ldots, (p_n, x_n)\}$. Then for some $p, C_{\text{total}} > 0$, we have for all $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$|L_{\text{total}}(\hat{\mu}_n) - L_{\text{total}}(\mu)| \leq C_{\text{total}} \left(\frac{\log(1/\delta)}{n} \right)^{1/p}. \tag{4}$$

Assumption 2 asserts that L_{total} behaves like a well-behaved statistical quantity that converges with more i.i.d. samples to the population loss. As such, we believe it is a natural assumption that ought to be satisfied by any reasonably-behaving loss one may consider. Indeed, we show Assumption 2 is necessary in Section C, and that it is satisfied by FID and FI in Example 1, Example 2. Note that when $\mathcal{X} = \mathbb{R}^d$, C_{total} may depend on d. Furthermore note $p = 2$ in Assumption 2 corresponds to sub-Gaussian concentration.

Assumption 3 (Convexity of L_{total})**.** *The loss L_{total} is convex in its argument: for any probability distributions μ_1, μ_2 and any $\lambda \in [0, 1]$,*

$$L_{\text{total}}(\lambda\mu_1 + (1 - \lambda)\mu_2) \leq \lambda L_{\text{total}}(\mu_1) + (1 - \lambda) L_{\text{total}}(\mu_2). \tag{5}$$

Assumption 3 is satisfied by a range of losses used in practice such as FID, FI, FVD, MMD, and many others. This is because these losses all can be written as convex combinations of functions which are convex in the space of distributions. We again show the necessity of Assumption 3 in Section C. Finally, we will impose the following Assumption 4 for technical convenience.

Assumption 4 (Boundedness of L_{total})**.** *The loss $L_{\text{total}} \in [0, B_{\text{total}}]$ for some $B_{\text{total}} > 0$.*

We remark that the per-round loss ℓ_{round} clearly satisfies Assumptions 2, 3, 4 if ℓ_{round} is bounded (or more generally if $\ell_{\text{round}}(\mu)$ is sub-Gaussian): Assumption 2 follows by Hoeffding's Inequality, Assumption 3 follows immediately, and Assumption 4 follows immediately. As canonical and concrete examples, we discuss how FID and FI satisfy Assumptions 2, 3, 4.

Example 1 (FID)**.** *Fréchet Inception Distance (FID) (Heusel et al., 2017b) is a standard method for evaluating image generation. It is defined by mapping generated images and real images to a latent feature space via the Inception network (Szegedy et al., 2015) and fitting the generated and real distributions in the latent spaces to Gaussians $\mathcal{N}(\theta, \Sigma), \mathcal{N}(\theta_{\text{ref}}, \Sigma_{\text{ref}})$ respectively. Suppose μ, μ_{ref} are supported in $\mathbb{B}(0, R)$, a ball of radius R centered at the origin; for example, this holds for μ_{ref} when μ_{ref} is ImageNet (Deng et al., 2009). Letting W_2 denote the 2-Wasserstein distance, FID is defined by*

$$\text{FID}(\mu, \mu_{\text{ref}}) := W_2^2(\mathcal{N}(\theta, \Sigma), \mathcal{N}(\theta_{\text{ref}}, \Sigma_{\text{ref}})) = \|\theta - \theta_{\text{ref}}\|_2^2 + \text{tr}\big(\Sigma + \Sigma_{\text{ref}} - 2(\Sigma\Sigma_{\text{ref}})^{1/2}\big). \tag{6}$$

Then, Assumption 2 is satisfied by Lemma 6, Assumption 3 is satisfied by Lemma 9, and Assumption 4 is satisfied by Lemma 10 with $B_{\text{total}} = O(R^2 d)$[3]. Thus $C_{\text{total}} = O(d^{5/2} R^4)$. These Lemmas are proven in Section F.

Example 2 (FI)**.** *The Frontier Integral (FI) loss, which is similar to the MAUVE loss used for text (Pillutla et al., 2021), was introduced in Liu et al. (2021). See (2) of Liu et al. (2021) for its definition. Assumption 2 is satisfied by Theorem 1 of Liu et al. (2021), Assumption 3 holds as FI is is a convex combination of Kullback-Leibler (KL) divergences between the outputted distribution and the target distribution, and Assumption 4 follows as FI is bounded in $[0, 1]$ as per page 3 of Liu et al. (2021). Here we use that since KL is convex (here the target distribution is not an argument as it is fixed).*

On global losses, technical challenges, and distinctions from the bandit literature: In contrast to traditional bandit problems, we must deal with challenging *global* losses such as FID, MAUVE, and many others used in practice. The key challenge is that such global scores do not afford a simple decomposition as a sum of round-by-round losses. For example, FID depends on the empirical mean and covariance of the entire corpus of generated data; thus the contribution of a single sample x_t

[3] R.e. Assumption 2, when $n < \log(1/\delta)$, the bound from Lemma 6 is > 1. Thus Assumption 2 follows by Assumption 4 after multiplying constants by B_{total}. When $n \geq \log(1/\delta)$, Assumption 2 follows immediately by Lemma 6.

generated on a single round t cannot be easily singled out. Dealing with these challenging global losses is the main technical innovation in our work.

At a high level, we decompose the global loss via Assumption 3, and then use the concentration assumption Assumption 2 to map to optimism. This enables the success of certain algorithmic principles such as UCB and Explore-then-Exploit under global losses. For more details, see the proof ideas of Theorem 1 in Section 3, presented after the Theorem. However, it is a-priori unclear that other bandit-style algorithms such as EXP3 (Auer et al., 1995) or EXP4 (Auer et al., 2002b) can succeed under these global losses.

Settings with global losses have been considered in the bandit and online learning literature, see Even-Dar et al. (2009; 2010); Rakhlin et al. (2011). All these papers work under full-information settings, in contrast to the partial-information feedback model considered here. For example, Even-Dar et al. (2009) (adversarial setting) and Even-Dar et al. (2010) (stochastic setting) assume a vector of losses over all k arms, which is then combined over all T rounds to create particular global losses. We note that our Assumption 2, Assumption 3, Assumption 4 on the global loss that we consider are noticeably less specific than the form of losses considered in Even-Dar et al. (2009; 2010).

Finally, we remark that as with many theoretically motivated decision-making algorithms, the parameters in our algorithms are stated using problem-dependent constants to facilitate analysis. In practice, these constants are typically treated as tunable hyperparameters; indeed in our experiments in Section 5, we found our algorithms to be robust to simple choices.

3 Routing to Compete with the Best Generative Model

We begin with the following setting, where we aim to develop an online routing algorithm that performs nearly as well as the best model m_i in hindsight. Consider the empirical measure on

$$\{(p_1, m_i(p_1)), \ldots, (p_T, m_i(p_T))\} := \hat{\mu}'_i \quad \text{for all} \quad i \in [K]. \tag{7}$$

Following the notation from Section 1, we define the regret Reg_T after T rounds of generation by

$$\mathrm{Reg}_T := L_{\mathrm{total}}(\hat{\mu}_T) - \min_{i \in [K]} L_{\mathrm{total}}(\hat{\mu}'_i). \tag{8}$$

The above notion of regret reflects that we want to perform nearly as well as the generative model with the best generation performance in hindsight, after all T rounds of generation are completed.

In this setting, we establish that the Upper Confidence Bound (UCB) algorithm, given in Algorithm 1, obtains a $\tilde{O}(T^{-1/p})$ regret bound. This is in the following Theorem 1, and proof ideas are discussed after the statement of the Theorem. Note that the prompts p_t are still used in Algorithm 1, as inputs by the chosen model m_{i_t} to produce the observed generations x_t. To set up notation, we consider the empirical measure on

$$\left\{\left(p_{t'}, m_{i_{t'}}(p_{t'})\right) : i_{t'} = i, 1 \le t' \le t\right\} := \hat{\mu}_{i,t} \quad \text{for all} \quad i \in [K]. \tag{9}$$

Algorithm 1 UCB Algorithm

1: **for** $t = 1$ to K **do**
2: Pick $i_t = t$
3: Observe generation $x_t = m_{i_t}(p_t)$
4: Set $\hat{L}_{i,K} = L_{\mathrm{total}}(\hat{\mu}_{i,K})$, $n_{i,K} = 1$ for all $i \in [K]$
5: **for** $t = K + 1$ to T **do**
6: Pick

$$i_t = \arg\min_{i \in [K]} \left\{\mathrm{UCB}_{i,t-1} := \hat{L}_{i,t-1} - \sigma(n_{i,t-1})\right\} \text{ where } \sigma(n_{i,t}) = C_{\mathrm{total}}\left(\log(2Kt^2)/n_{i,t}\right)^{1/p}$$

7: Update $n_{i_t,t} = n_{i_t,t-1} + 1$
8: Observe generation $x_t = m_{i_t}(p_t)$
9: Update $\hat{L}_{i,t} = L_{\mathrm{total}}(\hat{\mu}_{i,t})$

Theorem 1. *Suppose Assumptions 1, 2, 3, 4 hold and $p \geq 1$, where p is from Assumption 2. Then Algorithm 1 enjoys the following bound on expected regret: for any $T > K$, defining Reg_T as per (8),*

$$\mathbb{E}[\mathrm{Reg}_T] \leq O\Big(C_{\mathrm{total}}(p-1)^{\frac{1}{p} \vee \frac{1-p}{p}} \log(2KT)^{1/p} T^{-1/p} + B_{\mathrm{total}} K T^{-1}\Big).$$

In particular, when $p = 2$ in Assumption 2 (corresponding to sub-Gaussian concentration), we have

$$\mathbb{E}[\mathrm{Reg}_T] \leq O\big(C_{\mathrm{total}} \log(2KT)^{1/2} T^{-1/2} + B_{\mathrm{total}} K T^{-1}\big).$$

Theorem 1 is proven in Subsection A.1. At a high level, the proof of Theorem 1 uses Assumption 2, Assumption 3, Assumption 4 to map to optimism, the key idea behind UCB style algorithms. In particular, while we do not have a sum of per round loss that affords us linearity, the convexity in Assumption 3 gives us the first step of $L_{\mathrm{total}}(\hat{\mu}_T) \leq \sum_{i=1}^{K} (n_{i,T}/T)\, L_{\mathrm{total}}(\hat{\mu}_{i,T})$ to convert the global loss L_{total} to a weighted sum of model-by-model quantities. Then, we use the concentration assumption Assumption 2 to map to optimism. Namely, Assumption 2 shows the empirical value of L_{total} is a sufficiently accurate estimate of its true value with high enough probability. This enables us to then follow the traditional proof strategy of UCB (Lattimore & Szepesvári, 2020).

In Section C we complement the above Theorem by showing that if one only has concentration at say $p = 2$ without convexity, a lower bound of $1/T^{1/3}$ on regret bound is possible. This shows convexity is needed to get the promised rate in the upper bound and without it, at the very least one has a slower rate for regret. Similarly, we also show lower bounds on rate that match the upper bounds shown above, thus establishing that the rate one gets from the concentration Assumption 2 cannot be improved in general.

4 Routing to combine several lightweight, low-cost models

We now consider when we have a set of lightweight, low cost generative models with different strengths on different generation tasks, and a class $\mathcal{F}$ of routing functions mapping prompts to models. For example, each routing function in $\mathcal{F}$ maps prompts to the model $m_i, i \in [K]$ that best suits the associated generation task. Indeed, model routing is motivated by the prohibitively high inference cost of LLMs, and furthermore because different generative models have different strengths on different types of generation tasks, as discussed in Section 1. Routing functions in $\mathcal{F}$ represent prior knowledge on the strengths of different generative models for different generation tasks. Consequently we believe this setting is relevant for the practical applications of model routing.

We would like to design a cost-efficient and time-efficient algorithm that rather than routing the prompt to a single large, costly generative model, chooses one model among K lightweight models that best suits a given prompt. In hindsight after T rounds of generation, there is an optimal routing function $\hat{f}^\star \in \mathcal{F}$. Formally, defining the empirical measure

$$\big\{\big(p_1, m_{f(p_1)}(p_1)\big), \ldots, \big(p_T, m_{f(p_T)}(p_T)\big)\big\} := \hat{\mu}'_{f,T} \quad \text{for all} \quad f \in \mathcal{F}, \tag{10}$$

we have

$$f^\star = \operatorname*{argmin}_{f \in \mathcal{F}} L_{\mathrm{total}}\big(\hat{\mu}'_{f,T}\big). \tag{11}$$

Here, $f^\star$ represents the best way of routing the prompts to the lightweight models in hindsight, among the possibilities in $\mathcal{F}$. Of course, the prompts come in one-by-one and therefore $f^\star$ is not known until all T rounds of generation are complete. We would like to design an online routing algorithm that performs nearly as well as this optimal routing function $f^\star$.

We consequently define regret in this setting as follows:

$$\mathrm{Reg}_T := L_{\mathrm{total}}(\hat{\mu}_T) - \min_{f \in \mathcal{F}} L_{\mathrm{total}}\big(\hat{\mu}'_{f,T}\big) = L_{\mathrm{total}}(\hat{\mu}_T) - L_{\mathrm{total}}\big(\hat{\mu}'_{f^\star,T}\big). \tag{12}$$

Consider the following notion of an ERM oracle for this problem:

$$\mathrm{Oracle}_{\mathrm{ERM}}(\hat{\mu}_T) = \operatorname*{argmin}_{f \in \mathcal{F}} L_{\mathrm{total}}(\{(p_t, x_t) \text{ s.t. } i_t = f(p_t)\}). \tag{13}$$

In the natural manner, we can extend the definition of (13) to arbitrary empirical distributions arising in the learning procedure defined in Section 1.

4.1 Regret bound under general conditions

First, we show that we can achieve a $\tilde{O}(T^{-1/3})$ regret bound under Assumption 1, Assumption 2, Assumption 3, Assumption 4, using the Explore-then-Exploit algorithm. The algorithm is presented in Algorithm 2, and the corresponding regret bound is presented in Theorem 2.

Algorithm 2 Explore then Exploit

1: Set $C = \lfloor T^{\frac{p}{p+1}} \rfloor$, where $p \geq 1$ is from Assumption 2
2: **for** each routing function $f \in \mathcal{F}$ **do**
3: Create routing function generation set $S_f = \varnothing$
4: **for** $i = 1$ to K ('explore' phase) **do**
5: **for** $t = (i-1)C + 1$ to iC **do**
6: Observe prompt p_t
7: Set $i_t = i$
8: Observe generation $x_t = m_{i_t}(p_t)$
9: **for** each routing function $f \in \mathcal{F}$ **do**
10: **if** $f(p_t) = i_t$ **then**
11: Add x_t to S_f
12: Let $\hat{f}^\star = \operatorname{argmin}_{f \in \mathcal{F}} L_{\text{total}}(S_f)$
13: **for** $t = KC + 1$ to T ('exploit' phase) **do**
14: Set $i_t = \hat{f}^\star(p_t)$
15: Observe generation $x_t = m_{i_t}(p_t)$

Theorem 2. *Suppose Assumptions 1, 2, 3, 4 hold. Then, for an appropriate choice of C, with probability at least $1-\delta$, Algorithm 2 achieves a regret of*

$$\mathrm{Reg}_T \leq O\Big(\big\{B_{\text{total}}K + C_{\text{total}}\log(|\mathcal{F}|/\delta)^{1/p}\big\}T^{-\frac{1}{p+1}}\Big) \quad \textit{for} \quad T = \Omega\Big((K\log(|\mathcal{F}|/\delta))^{\frac{p+1}{p}}\Big).$$

Thus when $p = 2$ in Assumption 2 (sub-Gaussian concentration), with probability at least $1-\delta$,

$$\mathrm{Reg}_T \leq O\Big(\big\{B_{\text{total}}K + C_{\text{total}}\log(|\mathcal{F}|/\delta)^{1/2}\big\}T^{-\frac{1}{3}}\Big) \quad \textit{for} \quad T = \Omega\Big((K\log(|\mathcal{F}|/\delta))^{3/2}\Big).$$

Moreover, Algorithm 2 is ERM oracle efficient, where the definition of the ERM oracle is in (13).

We prove Theorem 2 in Section B.1. Note ERM oracle efficiency follows directly since Line 11 of Algorithm 2 can be directly implemented by calling that oracle on the true values of p_t, i_t and x_t.

4.2 Improving the regret bound: the power of choosing two models per round

We next consider a situation where we can query an additional generative model $m_{i'_t}$ and observe the additional generation $x'_t = m_{i'_t}(p_t)$ per round, as discussed in Remark 2. This is a natural situation in the context of Section 4, as our motivation in this Section is when the generative models are lightweight and low cost, and therefore querying an additional generative model can be done with minimal additional cost. Indeed, this section considers settings where the models being aggregated are significantly cheaper (perhaps an order of magnitude cheaper) than the state-of-the-art models that are compared in Section 3. In particular, the main concern of this section is the distribution of generated outputs displayed to the user; the additional generation x'_t is not displayed. In this situation, we establish that Algorithm 3 enjoys a $\tilde{O}\big(T^{-1/p}\big)$ regret bound, as stated in Theorem 3.

In Algorithm 3, we assume that the i'_t in addition to i_t are known when implementing (13). Note the history H_t comprises only of the prompts p_t and generations x'_t, but not the true generations x_t that are shown to the user. Our Theorem here now reads as follows:

Theorem 3. *Algorithm 3 achieves achieves the following bound on expected regret.*

$$\mathbb{E}[\mathrm{Reg}_T] \leq \frac{B_{\text{total}} \cdot \big(9K^2\log(2KT|\mathcal{F}|)^2 + 1\big)}{T} + C_{\text{total}}\left(\frac{\log(4T|\mathcal{F}|)}{T}\right)^{1/p}\left(1 + \frac{3(2K)^{1/p}}{2^{1-1/p}-1}\right). \quad (14)$$

We prove Theorem 3 in Section B.2.

Algorithm 3 Improving dependence on T

Initialize history $H_0 = \varnothing$
Initialize $m = 0$
for $t = 1$ to T **do**
  Observe prompt p_t
  if $t = 2^m$ **then**
    Increment m
    Set routing function $\hat{f}_m = \text{Oracle}_{\text{ERM}}(H_{t-1})$ (for $m = 0$, choose any $f \in \mathcal{F}$ arbitrarily)
  Sample a random model i'_t uniformly from $[K]$
  Observe generation $x'_t = m_{i'_t}(p_t)$
  Update history $H_t = H_{t-1} \cup \{(p_t, x'_t)\}$
  Select model $i_t = \hat{f}_m(p_t)$
  Observe generation $x_t = m_{i_t}(p_t)$

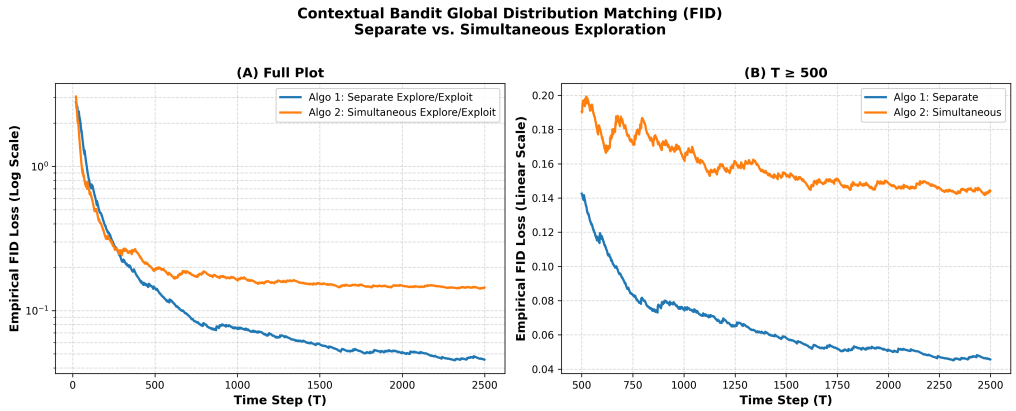

Figure 1: Regret of Algorithm 1 (Algorithm 6 from our work) vs Algorithm 2 (greedy algorithm)

5 Experiments

Here we empirically validate the conclusion of Section 4. We analyze a modification of Algorithm 3 in a topic-based setting (specifically, Algorithm 6) as described in Section D.1. In the topic-based setting, the intuition is that there may be different topics (e.g. math vs coding), and different models are better at each topic. The goal is to complete with the best model for each topic chosen separately. Our experiment uses a synthetic version of the problem with 4 experts, 4 topics, and $d = 8$ dimensions, where exactly one expert perfectly matches each topic (so the optimal policy has FID loss 0), and the other experts have shifted means. We provide details of the experiments in Section E.

We demonstrate that Algorithm 6, which uses an additional generation for exploration similarly to Algorithm 3, has vanishing FID loss as T grows large. In contrast, the naïve algorithm that simultaneously explores and exploits, similarly to Algorithm 3 but adapted to the topic setting (attempting a version of Algorithm 6 without an extra exploration step by greedily selecting the best model so far) has non-vanishing regret lower bounded by a constant as T grows large. These results are in Figure 1; (A) therein displays the regret over all rounds, while (B) displays the regret for rounds ≥ 500, which better captures the asymptotic regret behavior of both algorithms.

6 Conclusion

In this work, we introduced a general framework for online routing of generative models with both per-round and distributional losses, capturing practical objectives such as generation quality, diversity, and inference cost. We showed under suitable assumptions – many of which are necessary – variants of classical bandit algorithms can achieve vanishing regret. As such, our results provide a theoretical foundation for adaptive routing across generative models.

AI USE STATEMENT

In this work, we used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs and implementation, designing and providing feedback on research methodology or experiments, and implementing methods. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results. Additionally, we used generative AI tools for creating or editing software code, editing a research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. For example, we carefully read and checked all the paper's mathematical claims and proofs (by all authors), checked LLM-generated research ideas for potential plagiarism by manually reviewing the relevant literature (by all authors), checked all LLM-generated implementations and code for correctness (by 1 author), read and edited all LLM-written prose (by all authors), and carefully checked all LLM-suggested literature (by all authors). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors do not believe this paper raises any ethics questions.

REPRODUCIBILITY STATEMENT

All of the theorems discussed in this paper are fully reproducible due to the full proofs being provided in the appendices (Sections A to D and F). A significant number of experimental details are provided in Section E and full code for reproduction will be made available with publication.

REFERENCES

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. Jointly measuring diversity and quality in text generation models. In Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf (eds.), *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 90–98, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2311. URL https://aclanthology.org/W19-2311/.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pp. 322–331. IEEE, 1995.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

Sébastien Bubeck and Nicolo Cesa-Bianchi. *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, volume 5 of *Foundations and Trends in Machine Learning*. Now Publishers, 2014.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Kouider Chadli, Goetz Botterweck, and Takfarinas Saber. Pearl: Performance and energy aware routing for llms. *Future Generation Computer Systems*, pp. 108218, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chao-Kai Chiang, Takashi Ishida, and Masashi Sugiyama. Llm routing with dueling feedback. *arXiv preprint arXiv:2510.00841*, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Eyal Even-Dar, Robert Kleinberg, Shie Mannor, and Yishay Mansour. Online learning for global cost functions. In *COLT*, 2009.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Learning with global cost in stochastic environments. In *COLT*, pp. 80–92, 2010.

Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios N Angelopoulos, and Ion Stoica. Prompt-to-leaderboard. *arXiv preprint arXiv:2502.14855*, 2025.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of EMNLP*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017a.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017b. URL https://papers.nips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 0162-1459. doi: 10.2307/2282952. URL https://www.jstor.org/stable/2282952.

Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, April 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1843. URL https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-2/Mean-estimation-with-sub-Gaussian-rates-in-polynomial-time/10.1214/19-AOS1843.full.

Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. An optimism-based approach to online evaluation of generative models. *arXiv preprint arXiv:2406.07451*, 3, 2024.

Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. A multi-armed bandit approach to online selection and evaluation of generative models. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025a.

Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. Pak-ucb contextual bandit: An online learning approach to prompt-aware selection of generative models and llms. *nternational Conference on Machine Learning (ICML)*, 2025b.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, et al. Universal model routing for efficient llm inference. *arXiv preprint arXiv:2502.08773*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pp. 110–119, 2016.

Zhiqiu Lin et al. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024.

Lang Liu, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. Divergence frontiers for generative models: Sample complexity, quantization effects, and frontier integrals. *Advances in Neural Information Processing Systems*, 34:12930–12942, 2021.

Michael Luo, Justin Wong, Brandon Trabucco, Yanping Huang, Joseph E Gonzalez, Zhifeng Chen, Ruslan Salakhutdinov, and Ion Stoica. Stylus: Automatic adapter selection for diffusion models. *Advances in Neural Information Processing Systems*, 37:32888–32915, 2024.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Yasmin Moslem and John D Kelleher. Dynamic model routing and cascading for efficient llm inference: A survey. *arXiv preprint arXiv:2603.04445*, 2026.

Quang H Nguyen, Thinh Dao, Duy C Hoang, Juliette Decugis, Saurav Manchanda, Nitesh V Chawla, and Khoa D Doan. Metallm: A high-performant and cost-efficient dynamic framework for wrapping llms. *arXiv preprint arXiv:2407.10834*, 2024.

Victor M. Panaretos and Yoav Zemel. Fréchet Means in the Wasserstein Space $$\mathcal W_2$$. In Victor M. Panaretos and Yoav Zemel (eds.), *An Invitation to Statistics in Wasserstein Space*, pp. 59–74. Springer International Publishing, Cham, 2020. ISBN 978-3-030-38438-8. doi: 10.1007/978-3-030-38438-8_3. URL https://doi.org/10.1007/978-3-030-38438-8_3.

Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. Adaptive llm routing under budget constraints. *arXiv preprint arXiv:2508.21141*, 2025.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021.

Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. Diffusiongpt: Llm-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*, 2024.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Beyond regret. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 559–594. JMLR Workshop and Conference Proceedings, 2011.

Parham Rezaei, Farzan Farnia, and Cheuk Ting Li. Be more diverse than the most diverse: Optimal mixtures of generative models via mixture-ucb bandit algorithms. *arXiv preprint arXiv:2412.17622*, 2024.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2018.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. Efficient large language models: A survey. *Transactions on Machine Learning Research*, 2024.

Xinyuan Wang, Yanchi Liu, Wei Cheng, Xujiang Zhao, Zhengzhang Chen, Wenchao Yu, Yanjie Fu, and Haifeng Chen. Mixllm: Dynamic routing in mixed large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10912–10922, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Yu Xia, Fang Kong, Tong Yu, Liya Guo, Ryan A Rossi, Sungchul Kim, and Shuai Li. Which llm to play? convergence-aware online model selection with time-increasing bandits. In *Proceedings of the ACM Web Conference 2024*, pp. 4059–4070, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*, 2024.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *Proceedings of SIGIR*, pp. 1097–1100, 2018.

Thomas Ziller, Shashikant Ilager, Alessandro Tundo, Ezio Bartocci, Leonardo Mariani, and Ivona Brandic. Greenserv: Energy-efficient context-aware dynamic routing for multi-model llm inference. *arXiv preprint arXiv:2601.17551*, 2026.

A Omitted Proofs from Section 3

A.1 Proof of Theorem 1

Here we prove Theorem 1. The idea is that Assumption 3 lets us decompose Reg_T in a per-model fashion akin to the classical proof of UCB. By Lemma 9,

$$L_{\mathrm{total}}(\hat{\mu}_T) \le \sum_{i=1}^{K} \lambda_i L_{\mathrm{total}}(\hat{\mu}_{i,T}),$$

where $\lambda_i = n_{i,T}/T$ is the proportion of $t \in [T]$ such that $i_t = i$. Thus $\sum_{i=1}^{K} \lambda_i = 1$ and $\lambda_i \ge 0$ for all $i \in [K]$. We obtain

$$\mathrm{Reg}_T = L_{\mathrm{total}}(\hat{\mu}_T) - \min_{i\in[K]} L_{\mathrm{total}}(\hat{\mu}'_i) \le \sum_{i=1}^{K} \lambda_i L_{\mathrm{total}}(\hat{\mu}_{i,T}) - \min_{i\in[K]} L_{\mathrm{total}}(\hat{\mu}'_i). \tag{15}$$

Recalling the definition of μ_i in (3) and the definition of $\hat{L}_{i,t}$ in Algorithm 1, we define

$$L_i := L_{\mathrm{total}}(\mu_i), \tag{16}$$

$$i^\star := \operatorname*{argmin}_{i\in[K]} L_i, \tag{17}$$

$$\Delta_i := L_{\mathrm{total}}(\hat{\mu}_{i,T}) - L_{i^\star} = \hat{L}_{i,T} - L_{i^\star} \quad \text{for all} \quad i \in [K]. \tag{18}$$

$$\mathrm{err} := L_{i^\star} - \min_{i\in[K]} L_{\mathrm{total}}(\hat{\mu}'_i). \tag{19}$$

If $i^\star$ is not unique, we choose an arbitrary such index in $\operatorname{argmin}_{i\in[K]} L_i$. The motivation for the choice of $\Delta_i = L_{\mathrm{total}}(\hat{\mu}_{i,T}) - L_{i^\star}$ is given in the proof of Lemma 1 below. Also, notice $\mathrm{err} \ne 0$; $L_{i^\star}$ corresponds to *population* loss w.r.t. $\mu_{i^\star}$. Meanwhile, $\min_{i\in[K]} L_{\mathrm{total}}(\hat{\mu}'_i)$ is the empirical loss based on the best arm according to the particular set of prompts p_t that arrived.

Combining with (15), we have

$$\mathrm{Reg}_T \le \frac{1}{T} \sum_{i=1}^{K} n_{i,T} \Delta_i + \mathrm{err}. \tag{20}$$

We now are in a position to prove Theorem 1 via the classical proof strategy of UCB.

Lemma 1. *At any time $t \in [T]$ and any model $i \in [K]$,*

$$\Pr\left(s_{t+1} = i \,\middle|\, n_{i,t} \ge \left(\frac{2C_{\mathrm{total}}}{\Delta_i} \right)^p \log(2Kt^2) \right) \le t^{-2}.$$

Proof of Lemma 1. Observe that the (p_t, x_t) such that $i_t = i$ are drawn i.i.d. from μ_i. This is because the $p_t \sim \mathcal{D}_{\mathrm{prompts}}$ are i.i.d., since the choice of i_t in Algorithm 1 only depends on the past and does not depend on p_t, and since $x_t = m_i(p_t)$. This means each such $(p_t, x_t) \sim \mu_i$, and independence follows as the p_t are independent and i_t is adapted to the past. (This is the same rationale used in the classic proof of UCB (Lattimore & Szepesvári, 2020).) Consequently by Lemma 6, noting that at time t, model i is selected $n_{i,t}$ times, we obtain for all $\delta > 0$,

$$\Pr\left(\left| \hat{L}_{i,t} - L_i \right| \ge C_{\mathrm{total}} \left(\frac{\log(1/\delta)}{n_{i,t}} \right)^{1/p} \right) \le \delta\,.$$

In particular this applies when $t = T$ as well, giving a bound on $|\hat{L}_{i,T} - L_i|$. Now applying a Union Bound over $[K]$ and over rounds $\{t, T\}$, we conclude that for any $\delta > 0$ and all $i \in [K]$,

$$\delta \ge \Pr\big(\big|\hat{L}_{i,t} - L_i\big| \ge \sigma(n_{i,t})\big), \delta \ge \Pr\big(\big|\hat{L}_{i,T} - L_i\big| \ge \sigma(n_{i,T})\big) \ge \Pr\big(\big|\hat{L}_{i,T} - L_i\big| \ge \sigma(n_{i,t})\big), \tag{21}$$

where we used that $n_{i,t} \le n_{i,T}$, and where we define (note that this function has two arguments)

$$\sigma(n_{i,t}, \delta) := C_{\mathrm{total}} \left(\frac{\log(2K/\delta)}{n_{i,t}} \right)^{1/p} \ge 0.$$

Pick $\delta = 1/t^2$ in (21) and recall the definition $\sigma(n_{i,t}) := \sigma\big(n_{i,t}, 1/t^2\big) \geq 0$ in Algorithm 1. We obtain with probability at least $1 - 1/t^2$, for all $i \in [K]$,

$$|\hat{L}_{i,t} - L_i| \leq \sigma(n_{i,t}), |\hat{L}_{i,T} - L_i| \leq \sigma(n_{i,T}) \implies \mathrm{UCB}_{i,t} = \hat{L}_{i,t} - \sigma(n_{i,t}) \geq L_i - 2\sigma(n_{i,t}). \quad (22)$$

Observe that when $2\sigma(n_{i,t}) \leq \Delta_i$, under the event (22), we have

$$\begin{aligned} L_i - L_{i^\star} &\geq \hat{L}_{i,T} - \sigma(n_{i,t}) - L_{i^\star} \\ &= L_{\mathrm{total}}(\hat{\mu}_{i,t}) - L_{i^\star} - \sigma(n_{i,t}) \\ &= \Delta_i - \sigma(n_{i,t}) \\ &\geq 2\sigma(n_{i,t}), \end{aligned}$$

where we used that $\sigma(n_{i,T}) \leq \sigma(n_{i,t})$ as $n_{i,T} \leq n_{i,t}$ and as $\sigma(\cdot)$ is decreasing. Therefore

$$\begin{aligned} \mathrm{UCB}_{i,t} &= \hat{L}_{i,t} - \sigma(n_{i,t}) \\ &\geq L_i - 2\sigma(n_{i,t}) \\ &\geq L_{i^\star} \\ &\geq \hat{L}_{i^\star,t} - \sigma(n_{i^\star,t}) = \mathrm{UCB}_{i^\star,t}, \end{aligned}$$

where we used that under the event (21), $L_{i^\star} - \hat{L}_{i^\star,t} \geq -\sigma(n_{i^\star,t}), \hat{L}_{i,t} - L_i \geq -\sigma(n_{i,t})$.

Hence if $2\sigma(n_{i,t}) \leq \Delta_i$, with probability at least $1 - t^{-2}$, Algorithm 1 will not pick model i. Noting that

$$n_{i,t} \geq \left(\frac{2C_{\mathrm{total}}}{\Delta_i}\right)^p \log(2Kt^2) \implies 2\sigma(n_{i,t}) \leq \Delta_i,$$

the Lemma follows. □

Lemma 2. *For all models $i \in [K]$, we have that*

$$\mathbb{E}[n_{i,T}] \leq 2\left(\frac{2C_{\mathrm{total}}}{\Delta_i}\right)^p \log(2KT) + 8.$$

Proof of Lemma 2. This follows directly from Lemma 1. For this proof only, let us define

$$C_t := \max\left(\frac{2C_{\mathrm{total}}}{\Delta_i}\right)^p \log(2Kt^2).$$

Note that by definition of Algorithm 1,

$$\begin{aligned} \mathbb{E}[n_{i,t}] &= 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{i_t = i\}\right] \\ &= 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{i_t = i, n_{i,t} < C_t\}\right] + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{i_t = i, n_{i,t} \geq C_t\}\right] \\ &= 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{i_t = i, n_{i,t} < C_t\}\right] + \sum_{t=K+1}^{T} \Pr(i_t = i, n_{i,t} \geq C_t) \\ &= 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{i_t = i, n_{i,t} < C_t\}\right] + \sum_{t=K+1}^{T} \Pr(i_t = i \,|\, n_{i,t} \geq C_t) \\ &\leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{i_t = i, n_{i,t} < C_t\}\right] + \sum_{t=K+1}^{T} t^{-2} \\ &\leq 8 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{i_t = i, n_{i,t} < C_T\}\right], \end{aligned}$$

where we applied Lemma 1 and then bounded the resulting series. Now note that

$$\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{i_t = i, n_{i,t} < C_T\}\right] \leq C_T,$$

because if there are at least C_T indices t such that $\mathbf{1}\{i_t = i, n_{i,t} < C_T\} = 1$, letting t' be the C_T-th such index, we must have $n_{i,t''} \geq C_T$, and thus the indices $t'' \geq t' + 1$ do not contribute any further to the sum. Combining with the earlier display proves the Lemma. □

Finally, we will control err as follows:

Lemma 3. *We have*

$$\mathbb{E}[\mathrm{err}] \le O\left(C_{\mathrm{total}}\left(\frac{\log(1/\delta)}{T}\right)^{1/p}\right).$$

Proof of Lemma 3. Notice that

$$\min_{i\in[K]} L_{\mathrm{total}}(\hat{\mu}_i') \le L_{\mathrm{total}}(\hat{\mu}_{i^\star}').$$

Therefore by Lemma 6 and using the definition of $\mathcal{D}_{i^\star}$,

$$\begin{aligned}\mathbb{E}[\mathrm{err}] &= \mathbb{E}[L_{i^\star}] - \mathbb{E}\left[\min_{i\in[K]} L_{\mathrm{total}}(\hat{\mu}_i')\right] \\ &\le \mathbb{E}[L_{\mathrm{total}}(\mu_{i^\star}) - L_{\mathrm{total}}(\hat{\mu}_{i^\star}')] \\ &\le O\left(C_{\mathrm{total}}\left(\frac{\log(1/\delta)}{T}\right)^{1/p}\right).\end{aligned}$$

Here, the last inequality follows as all the $p_t \sim \mathcal{D}_{\mathrm{prompts}}$ are i.i.d., $\{m_{i^\star}(p_t) : t \in [T]\}$ are T i.i.d. draws from $\mu_{i^\star}$, upper bounding $L_{\mathrm{total}}(\mu_{i^\star}) - L_{\mathrm{total}}(\hat{\mu}_{i^\star}') \le |L_{\mathrm{total}}(\mu_{i^\star}) - L_{\mathrm{total}}(\hat{\mu}_{i^\star}')|$, and integrating the tail bound on $|L_{\mathrm{total}}(\mu_{i^\star}) - L_{\mathrm{total}}(\hat{\mu}_{i^\star}')|$ from Lemma 6. □

Finally, combining Lemma 2, Lemma 3, and (20) proves Theorem 1. In particular, by these results, it remains to upper bound $\mathbb{E}\left[\sum_{i=1}^K n_{i,T}\Delta_i\right]$. For any $C > 0$,

$$\mathcal{S}_1 := \left\{i : \Delta_i < \left(\frac{C}{T}\right)^{1/p}\right\}, \mathcal{S}_2 := \left\{i : \Delta_i \ge \left(\frac{C}{T}\right)^{1/p}\right\}.$$

Later, we will optimize over the constant $C > 0$.

Notice $\sum_{i\in\mathcal{S}_1} \mathbb{E}[n_{i,T}] = \mathbb{E}\left[\sum_{i\in\mathcal{S}_1} n_{i,T}\right] \le T$. It follows that

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^K n_{i,T}\Delta_i\right] &\le \left(\frac{C}{T}\right)^{1/p} \sum_{i\in\mathcal{S}_1} \mathbb{E}[n_{i,T}] + \sum_{i\in\mathcal{S}_2} \mathbb{E}[n_{i,T}]\Delta_i \\ &\le C^{1/p}T^{1-1/p} + \sum_{i\in\mathcal{S}_2} \mathbb{E}[n_{i,T}]\Delta_i.\end{aligned}$$

Note $\Delta_i \le 2B_{\mathrm{total}}$ by Assumption 4. By Lemma 2, we thus bound

$$\begin{aligned}\sum_{i\in\mathcal{S}_2} \mathbb{E}[n_{i,T}]\Delta_i &\le \sum_{i\in\mathcal{S}_2}\left(2\left(\frac{2C_{\mathrm{total}}}{\Delta_i}\right)^p \log(2KT) + 8\right)\Delta_i \\ &\le O\left((2C_{\mathrm{total}})^p \log(2KT)(C/T)^{\frac{1-p}{p}}\right) + O(B_{\mathrm{total}}K).\end{aligned}$$

Combining with (20) and Lemma 3 gives

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_T] &\le \mathbb{E}\left[\frac{1}{T}\sum_{i=1}^K n_{i,T}\Delta_i + \mathrm{err}\right] \\ &\le \frac{1}{T}\left\{C^{1/p}T^{1-1/p} + (2C_{\mathrm{total}})^p C^{\frac{1-p}{p}} \log(2KT)T^{1-1/p} + O(B_{\mathrm{total}}K)\right\} \\ &\le O\left(\left(C^{1/p} + 2C_{\mathrm{total}}\right)^p C^{\frac{1-p}{p}} \log(2KT)\right)T^{-1/p} + \frac{B_{\mathrm{total}}K}{T}\Big).\end{aligned}$$

Choosing

$$C = (p-1)(2C_{\mathrm{total}})^p \log(2KT),$$

we complete the proof of Theorem 1.

B Omitted Proofs from Section 4

B.1 Proof of Theorem 2

Preliminaries: First, we show that with high probability, every routing function accumulates a sufficient number of samples in its generation set S_f. Recall the definition of the S_f and also the choice

$$C = \lfloor T^{\frac{p}{p+1}} \rfloor \tag{23}$$

from Algorithm 2. For any fixed routing function $f \in \mathcal{F}$, we define the empirical measure on

$$\{(p_t, m_{f(p_t)}(p_t)) : t \le C\} := \hat{\mu}_f^{\text{explore}}. \tag{24}$$

Note we have

$$|S_f| = \sum_{i=1}^{K} \sum_{t=(i-1)C+1}^{iC} \mathbf{1}\{f(p_t) = i_t\}.$$

Recall that $i_t = i$ for $t \in [(i-1)C+1, iC]$. Since the prompts p_t are drawn i.i.d., the expected size of S_f is

$$\mathbb{E}[|S_f|] = \sum_{i=1}^{K} \sum_{t=(i-1)C+1}^{iC} \Pr(f(p_t) = i) = C \sum_{i=1}^{K} \Pr(f(p_t) = i) = C.$$

Notice that S_f is the sum of KC independent indicator variables. By Hoeffding's Inequality, we have

$$\Pr\left(|S_f| \le C - \sqrt{\frac{KC}{2} \log(3|\mathcal{F}|/\delta)}\right) \le \exp\left(-2 \cdot \frac{KC}{2} \log(3|\mathcal{F}|/\delta) / KC\right) = \frac{\delta}{3|\mathcal{F}|}.$$

Applying a Union Bound over all routing functions in $\mathcal{F}$, we conclude that with probability at least $1 - \delta/3$, every routing function $f \in \mathcal{F}$ satisfies

$$|S_f| \ge C' := C - \sqrt{\frac{KC}{2} \log(3|\mathcal{F}|/\delta)}. \tag{25}$$

Next, we argue that the empirical value of L_{total} for each routing function evaluated on S_f is close to its true loss. Let μ_f denote the true distribution induced by routing function f, formally defined as

$$\mu_f := (p, m_{f(p)}(p)) \quad \text{where} \quad p \sim \mathcal{D}_{\text{prompts}}. \tag{26}$$

Observe that by construction S_f comprises of $|S_f|$ i.i.d. samples from μ_f. Now by Assumption 2 with $n = |S_f|$ and failure probability $\frac{\delta}{3|\mathcal{F}|}$, we obtain that with probability at least $1 - \frac{\delta}{3|\mathcal{F}|}$,

$$\left|L_{\text{total}}(\hat{\mu}_f^{\text{explore}}) - L_{\text{total}}(\mu_f)\right| \le C_{\text{total}} \left(\frac{\log(6|\mathcal{F}|/\delta)}{|S_f|}\right)^{1/p}, \tag{27}$$

where we recall the definition of $\hat{\mu}_f^{\text{explore}}$ from (24). We now take a Union Bound over (25), (27) to obtain that with probability at least $1 - 2\delta/3$, we have for all $f \in \mathcal{F}$,

$$\left|L_{\text{total}}(\hat{\mu}_f^{\text{explore}}) - L_{\text{total}}(\mu_f)\right| \le \varepsilon_{\text{est}} \quad \text{where} \quad \varepsilon_{\text{est}} := C_{\text{total}} \left(\frac{\log(6|\mathcal{F}|/\delta)}{C'}\right)^{1/p}. \tag{28}$$

We will use (28) to upper bound the regret from the 'exploit' phase of Algorithm 2.

Bounding the overall regret: By definition, the empirical distribution $\{x_1, \ldots, x_T\}$ is a mixture of the empirical distribution from the exploration phase $\hat{\mu}_{\text{explore}}$ (of size KC) and the empirical distribution from the exploitation phase $\hat{\mu}_{\text{exploit}}$ (of size $T - KC$). For any fixed routing function $f \in \mathcal{F}$, we define the empirical measure on

$$\{(p_t, m_{f(p_t)}(p_t)) : t > C\} := \hat{\mu}_f^{\text{exploit}}. \tag{29}$$

Notice that $\hat{\mu}_{\text{exploit}} \equiv \hat{\mu}^{\text{exploit}}_{\hat{f}^\star}$. Therefore by Assumption 3, we have

$$\begin{aligned} L_{\text{total}}(\hat{\mu}_T) &\le \frac{KC}{T} L_{\text{total}}(\hat{\mu}_{\text{explore}}) + \frac{T-KC}{T} L_{\text{total}}(\hat{\mu}_{\text{exploit}}) \\ &= \frac{KC}{T} L_{\text{total}}(\hat{\mu}_{\text{explore}}) + \frac{T-KC}{T} L_{\text{total}}(\hat{\mu}^{\text{exploit}}_{\hat{f}^\star}). \end{aligned} \tag{30}$$

We upper bound both the terms in (30) as follows:

1. By Assumption 4 and as $C = \lfloor T^{\frac{p}{p+1}} \rfloor$, we have
$$\frac{KC}{T} L_{\text{total}}(\hat{\mu}_{\text{explore}}) \le B_{\text{total}} K T^{-\frac{1}{p+1}}. \tag{31}$$

2. By definition of $\hat{f}^\star$, we have
$$L_{\text{total}}(\hat{\mu}^{\text{explore}}_{\hat{f}^\star}) \le L_{\text{total}}(\hat{\mu}^{\text{explore}}_{f^\star}).$$
Combining with (28) gives with probability at least $1 - 4\delta/3$,
$$\begin{aligned} L_{\text{total}}(\mu_{\hat{f}^\star}) &\le L_{\text{total}}(\hat{\mu}^{\text{explore}}_{\hat{f}^\star}) + \varepsilon_{\text{est}} \\ &\le L_{\text{total}}(\hat{\mu}^{\text{explore}}_{f^\star}) + \varepsilon_{\text{est}} \\ &\le L_{\text{total}}(\mu_{f^\star}) + 2\varepsilon_{\text{est}}. \end{aligned} \tag{32}$$
Note for any $f \in \mathcal{F}$ that $(p_t, m_{f(p_t)}(p_t))$ are i.i.d. from μ_f by Assumption 1. By Assumption 2 and our lower bound in T as a condition of the Theorem, we have with probability at least $1 - \delta/3$,
$$\left| L_{\text{total}}(\hat{\mu}^{\text{exploit}}_{\hat{f}^\star}) - L_{\text{total}}(\mu_{\hat{f}^\star}) \right| \le \varepsilon_{\text{est}} \quad , \quad \left| L_{\text{total}}(\hat{\mu}'_{f^\star,T}) - L_{\text{total}}(\mu_{f^\star}) \right| \le \varepsilon_{\text{est}}.$$
Here we recall the definition of $f^\star$ in (11) and $\hat{\mu}'_{f^\star,T}$ from (10). Combining with (32) gives with probability at least $1 - 2\delta$,
$$L_{\text{total}}(\hat{\mu}^{\text{exploit}}_{\hat{f}^\star}) - L_{\text{total}}(\hat{\mu}'_{f^\star,T}) \le 4\varepsilon_{\text{est}}.$$

Combining the above cases with (30) gives with probability at least $1 - 2\delta$,

$$\text{Reg}_T \le B_{\text{total}} K T^{-\frac{1}{p+1}} + L_{\text{total}}(\hat{\mu}^{\text{exploit}}_{\hat{f}^\star}) - L_{\text{total}}(\hat{\mu}_{f^\star,T}) \le B_{\text{total}} K T^{-\frac{1}{p+1}} + 4\varepsilon_{\text{est}},$$

where we again recall the definition of $f^\star$ in (11). Recalling the choice of C, the definition of ε_{est}, and our lower bound in T as a condition of the Theorem and rearranging the above bound, the desired regret bound follows.

B.2 Proof of Theorem 3

We will assume that the number of rounds T is of the form $2^k - 1$ (for $k \in \mathbb{N}$) for the sake of clarity, but note that this assumption could easily be removed.

As before, let μ_f denote the true distribution induced by routing function f, formally defined as

$$\mu_f := \big(p, m_{f(p)}(p)\big) \quad \text{where} \quad p \sim \mathcal{D}_{\text{prompts}}. \tag{33}$$

and let $\hat{\mu}_{f,T}$ be the empirical distribution over the T rounds, formally

$$\mu_f := \big(p, m_{f(p)}(p)\big) \quad \text{where} \quad p \sim \{p_t : 1 \le t \le T\}. \tag{34}$$

Note that the there are $\log_2(T)$ different ERM oracles used throughout the T steps, due to the exponential epoch schedule.

For all $m, 1 \le m \le \log_2(T)$, the number of rounds that $\hat{f}_m$ is played as an ERM oracle is 2^{m-1}.

Denote the following empirical measures, where x_t, x'_t are defined as in Algorithm 3:

$$\hat{\mu}_m \coloneqq \left\{(p_t, x_t) \,:\, 2^{m-1} \le t < 2^m - 1\right\} \quad \text{for all} \quad m \in [\log_2(T)],$$
$$\hat{\mu}^{\text{history}}_{f,t} \coloneqq \left\{(p_{t'}, x'_{t'}) \,:\, \hat{f}_{\lfloor \log_2(t')+1 \rfloor}(p_t) = f(p_t), 1 \le t' \le t\right\} \quad \text{for all} \quad f \in \mathcal{F},$$
$$\Delta_m \coloneqq \hat{\mu}^{\text{history}}_{f,T} - L_{\text{total}}(\hat{\mu}_{f^\star,T}).$$

where $f^* = \arg\min_f L_{\text{total}}(\hat{\mu}_{f,T})$. We observe that by Assumption 3,

$$\text{Reg}_T \le \sum_{m=1}^{\log_2(T)} \frac{2^{m-1}}{T}(L_{\text{total}}(\hat{\mu}_m) - L_{\text{total}}(\hat{\mu}_{f^\star,T}))$$

Now, by Assumption 2, with probability at least $1 - \delta/4$,

$$|L_{\text{total}}(\hat{\mu}_m) - L_{\text{total}}(\mu_{f_m})| \le C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{n_m}\right)^{1/p}$$

where n_m is defined as the cardinality of $\{\hat{f}_m(p_t) = x'_t \,:\, 1 \le t < 2^m\}$, and we have taken a union bound.

Similarly, with probability at least $1 - \delta/4$,

$$\left|L_{\text{total}}(\hat{\mu}^{\text{history}}_{f_m,2^m-1}) - L_{\text{total}}(\mu_{f_m})\right| \le C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{2^m - 1}\right)^{1/p}$$

In addition, by Assumption 2, with probability at least $1 - \delta/4$,

$$|L_{\text{total}}(\hat{\mu}_{f^\star,T}) - L_{\text{total}}(\mu_{f^\star})| \le C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{T}\right)^{1/p}$$

Similarly, with probability at least $1 - \delta/4$,

$$\left|L_{\text{total}}(\hat{\mu}^{\text{history}}_{f^\star,2^m-1}) - L_{\text{total}}(\mu_{f^\star})\right| \le C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{2^m - 1}\right)^{1/p}$$

Taking a union bound over the above 4 assumptions of Assumption 2, all of them hold with probability at least $1 - \delta$. In this case, we have

$$L_{\text{total}}(\hat{\mu}_m) - L_{\text{total}}(\hat{\mu}_{f^\star,T}) \le C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{2^m - 1}\right)^{1/p} + C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{n_m}\right)^{1/p}$$
$$+L_{\text{total}}(\hat{\mu}^{\text{history}}_{f_m,2^m-1}) - L_{\text{total}}(\hat{\mu}^{\text{history}}_{f^\star,2^m-1}) + C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{T}\right)^{1/p} + C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{2^m - 1}\right)^{1/p}$$

In addition, by the definition of the ERM, we have that

$$L_{\text{total}}(\hat{\mu}^{\text{history}}_{f_m,2^m-1}) - L_{\text{total}}(\hat{\mu}^{\text{history}}_{f^\star,2^m-1}) \le 0$$

Therefore,

$$L_{\text{total}}(\hat{\mu}_m) - L_{\text{total}}(\hat{\mu}_{f^\star,T}) \le C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{2^m - 1}\right)^{1/p} + C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{n_m}\right)^{1/p} + C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{T}\right)^{1/p}$$
$$+C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{2^m - 1}\right)^{1/p}$$

By Hoeffding's inequality, we have that

$$n_m = C(m) - \sqrt{\frac{KC(m)}{2}\log\left(\frac{2Kt^2|\mathcal{F}|}{\delta}\right)} \text{ with } C(m) = \frac{2^m - 1}{K}$$

as in the proof of Theorem 2.

Now, suppose that $2^m - 1 > 9K^2\log(2K\mathcal{F}/\delta)^2$. We will show that $C' \geq \frac{C}{2}$. We have that

$$\begin{aligned}
\sqrt{\frac{KC}{2}\log\left(\frac{2Kt^2|\mathcal{F}|}{\delta}\right)} &= \sqrt{\frac{t}{2K}\log\left(\frac{2Kt^2|\mathcal{F}|}{\delta}\right)} \\
&\leq \sqrt{\frac{t}{2K}\cdot 4\log(\sqrt{t})\cdot\log\left(\frac{2K|\mathcal{F}|}{\delta}\right)} \\
&\leq \sqrt{\frac{t}{2K}\cdot\frac{4\sqrt{t}}{e}\cdot\log\left(\frac{2K|\mathcal{F}|}{\delta}\right)} \\
&\leq \sqrt{\frac{t}{2K}\cdot\frac{4\sqrt{t}}{e}\cdot\sqrt{\frac{t}{9K^2}}} \\
&\leq \sqrt{\frac{t}{2K}\cdot\frac{t}{2K}} \\
&= \frac{t}{2K} = \frac{C}{2},
\end{aligned}$$

as desired. Now, for $t > 9K^2\log(2K|\mathcal{F}|/\delta)^2$, we then have

$$L_{\text{total}}(\hat{\mu}_m) - L_{\text{total}}(\hat{\mu}_{f^\star,T}) \leq C_{\text{total}}\left(\frac{\log(4\mathcal{F}/\delta)}{T}\right)^{1/p} + 3C_{\text{total}}\left(\frac{2\log(4\mathcal{F}/\delta)}{2^m-1}\right)^{1/p}$$

Substituting this upper bound into the regret decomposition, and using the trivial upper bound of B_{total} for the epochs where $2^m - 1 \leq 9K^2\log(2K|\mathcal{F}|/\delta)^2$, gives with probability at least $1-\delta$,

$$\begin{aligned}
\text{Reg}_T &\leq \sum_{m=1}^{\log_2(T)}\frac{2^{m-1}}{T}\left(C_{\text{total}}\left(\frac{\log(4|\mathcal{F}|/\delta)}{T}\right)^{1/p} + 3C_{\text{total}}\left(\frac{2K\log(4|\mathcal{F}|/\delta)}{2^m-1}\right)^{1/p}\right) \\
&\quad + \frac{B_{\text{total}}\cdot 9K^2\log(2K|\mathcal{F}|/\delta)^2}{T} \leq \frac{B_{\text{total}}\cdot 9K^2\log(2K|\mathcal{F}|/\delta)^2}{T} + C_{\text{total}}\left(\frac{\log(4|\mathcal{F}|/\delta)}{T}\right)^{1/p} \\
&\quad + \frac{3C_{\text{total}}(2K\log(4|\mathcal{F}|/\delta))^{1/p}}{T}\sum_{m=1}^{\log_2(T)}\frac{2^{m-1}}{(2^m-1)^{1/p}}.
\end{aligned}$$

Using the fact that $\frac{2^{m-1}}{(2^m-1)^{1/p}} \leq 2^{(m-1)(1-1/p)}$, we can upper bound the summation by a geometric series. For $p > 1$, we obtain

$$\sum_{m=1}^{\log_2(T)} 2^{(m-1)(1-1/p)} \leq \frac{T^{1-1/p}}{2^{1-1/p}-1}.$$

To obtain an expected regret bound, we set the failure probability to $\delta = 1/T$. In the failure event, the maximum possible value of the average regret Reg_T is trivially bounded by B_{total}. Thus, the expected average regret is bounded by

$$\begin{aligned}
\mathbb{E}[\text{Reg}_T] &\leq \left(\frac{B_{\text{total}}\cdot 9K^2\log(2KT|\mathcal{F}|)^2}{T} + C_{\text{total}}\left(\frac{\log(4T|\mathcal{F}|)}{T}\right)^{1/p}\left(1 + \frac{3(2K)^{1/p}}{2^{1-1/p}-1}\right)\right) + \frac{B_{\text{total}}}{T} \\
&= \frac{B_{\text{total}}\cdot\left(9K^2\log(2KT|\mathcal{F}|)^2 + 1\right)}{T} + C_{\text{total}}\left(\frac{\log(4T|\mathcal{F}|)}{T}\right)^{1/p}\left(1 + \frac{3(2K)^{1/p}}{2^{1-1/p}-1}\right).
\end{aligned}$$

C Lower Bounds

In this section we show provide lower bounds to show that some of the Assumptions made in the paper are required, and also argue that some of the rates on regret we claim are right. Note the dependence on the number of models in the regret bound of Algorithm 1 is tight up to log factors with respect to uniform bounds that apply to all global loss functions, as seen by the lower bound in the traditional per-round loss multi-armed bandit setting.

We first start by showing not only that some form of concentration of total loss is required, but in fact showing that the rate obtained from such a concentration bound (Assumption 2) cannot really be improved. In particular, we show that the dependence on the constants p and C_{total} from Assumption 2 are in some sense optimal. Consider a distribution μ with $\mathcal{P} = \{\}$ and some constants $p, C_{\text{total}}, \delta > 0$ such that for any $n \in \mathbb{N}$, we have with probability at least $c/2$ both

$$L_{\text{total}}(\hat{\mu}_n) - L_{\text{total}}(\mu) \geq \frac{1}{2} c\, C_{\text{total}} n^{-1/p} \text{ and } L_{\text{total}}(\hat{\mu}_n) - L_{\text{total}}(\mu) \leq -\frac{1}{2} c\, C_{\text{total}} n^{-1/p}, \tag{35}$$

where $\hat{\mu}_n$ is the empirical distribution of n i.i.d. samples from μ. Here, the condition $\mathcal{P} = \{\}$ simply means that L_{total} does not depend on the prompt. For example, the FID loss satisfies (35) with $p = 2$, where C_{total} is sufficiently small in terms of μ. We now have the following Theorem 4.

Theorem 4 (Concentration Lower Bound)**.** *For any distribution μ satisfying (35), there exists generative models m_1, m_2 such that the following holds. For any algorithm $\mathcal{A}$, letting Reg_T be the regret corresponding to $\mathcal{A}$ as per (8), we have*

$$\mathbb{E}[\text{Reg}_T] \geq \frac{1}{2} c^2\, C_{\text{total}} T^{-1/p}, \tag{36}$$

where C_{total}, c, and p are from (35).

Proof. Let $\mu_{\mathcal{X}}$ be the marginal over generation set $\mathcal{X}$ of μ. We construct the generative models m_1, m_2 as follows: say m_1, m_2 each an output in $\mathcal{X}$ drawn according to $\mu_{\mathcal{X}}$, using independent sources of randomness. Then, from the perspective of algorithm $\mathcal{A}$, on any given round, the two models are indistinguishable as their outputs have same law. Thus we can assume, without loss of generality, that $\mathcal{A}$ always chooses m_1.

Now, by (35), with probability at least $c/2$, the loss of m_1 over the empirical distribution μ_n exceeds the loss of m_1 over μ by at least $C_{\text{total}}\left(\frac{1}{T}\right)^{1/p}$. Similarly, with probability at least $c/2$, the loss of m_2 over the empirical distribution μ_n exceeds the loss of m_2 over μ by at least $C_{\text{total}}\left(\frac{1}{T}\right)^{1/p}$.

Since the models generate outputs using independent sources of randomness, the two events described above are independent, and the probability that both occur simultaneously is at least $c^2/4$. Therefore, the expected regret of the algorithm is at least

$$\frac{c^2}{4} \cdot 2 \cdot C_{\text{total}} \left(\frac{1}{T}\right)^{1/p} = \frac{C_{\text{total}} c^2}{2} \left(\frac{1}{T}\right)^{1/p}. \tag{37}$$

□

Next, we establish that the convexity assumption made in Assumption 3 is indeed required.

Theorem 5 (Convexity is Required)**.** *There exists a global loss L_{total} satisfying Assumption 2 with $p \leq 2$ but not satisfying Assumption 3, and a setting with just two models (i.e. $K = 2$), such that for any algorithm $\mathcal{A}$, defining Reg_T as the regret corresponding to $\mathcal{A}$ as per (8), $\text{Reg}_T = \Omega(T^{-1/3})$.*

Proof. Consider the case where $\mathcal{P} = \{\}$ and model $m_1(\cdot)$ samples from a distribution with mean $+1$ and standard deviation 1 and model $m_2(\cdot)$ samples from a distribution with mean -1 and standard deviation $1 - \varepsilon$. Now say $L_{\text{total}}(\mu) = \sqrt{\text{Var}(\mu)}$. In this case, for bounded variables one can easily see that the standard deviation concentrates at a rate of $1/\sqrt{n}$. Hence the concentration assumption Assumption 2 applies with $p = 2$. However, note that if we do not identify the right distribution (m_2, which has smaller standard deviation of $1 - \varepsilon$) then we pay a regret of ε. But then to identify the distribution with smaller variance (smaller by ε) we need at least $1/\varepsilon^2$ tries of samples from the

sub-optimal model m_1. Hence if we consider the case where we generated from model m_1 for $1/\varepsilon^2$ number of times out of T, then the standard deviation is lower bounded by $1-\varepsilon+\frac{1}{T\varepsilon^2}$. Hence regret is lower bounded by $1/T\varepsilon^2$. Now since the regret is lower bounded by the maximum between ε and $\frac{1}{T\varepsilon^2}$, optimizing over ε yields $\varepsilon = 1/T^{1/3}$, which is the claimed lower bound on regret. □

Finally, one might wonder if Assumption 1 – the assumption that prompts are drawn i.i.d. from a fixed distribution $\mathcal{D}_{\text{prompts}}$ – can be greatly relaxed. Towards answering this question, we show Theorem 6: that there is a problem setting with FID loss such that for any algorithm, with adversarial prompt generation in terms of the result of the algorithm, it is not be possible for the algorithm to achieve non-trivial regret. This represents a significant difference between the per-round loss setting studied in classic bandit literature and the global loss setting we consider for generation problems here.

Theorem 6 (Adversarial Setting Lower Bound)**.** *For any algorithm $\mathcal{A}$, suppose the prompts p_t are chosen adversarially in terms of $\mathcal{A}$. Then there exists a distribution μ_{ref} and generative models m_1, m_2 such that the following holds. Letting L_{total} be the FID loss, defining Reg_T as the regret corresponding to $\mathcal{A}$ as per (8), we have $\text{Reg}_T = \Omega(1)$.*

Proof. We let $\mathcal{X} = \mathbb{R}^d$ and $\mu_{\text{ref}} = \mathcal{N}(1/2, 1/4)$. Suppose that T is even, the proof is analogous for T odd. We construct the prompts and models as follows. Let $\mathcal{P} = \{0, 1, 2\}$, and construct the models by

1. Model 1: We let
$$m_1(0) = 0, m_1(1) = 1, m_1(2) = 0.$$
2. Model 2: We let
$$m_2(0) = 0, m_2(1) = 1, m_2(2) = 1.$$

We then construct the prompts as follows:

- We let $p_t = 2$ for all $1 \le t \le T/2$ (thus $m_1(p_t) = 0, m_2(p_t) = 1$ for all such t).
- We let $p_t = 1$ for all $t > T/2$ if m_1 was chosen less in the first $T/2$ rounds and $p_t = 0$ for all $t > T/2$ otherwise (and break ties arbitrarily).

Note m_1, m_2 now satisfy the following property:

- For the first $T/2$ rounds, m_1 will always output 0 and m_2 will always output 1.
- In the remaining $T/2$ rounds, both models output the value of the model m' which was chosen the fewest times in the first $T/2$ rounds.

Note in the first $T/2$ rounds, m' must have been chosen at most $T/4$ times.

Now, let r be the fraction of times that $x_t = 1$ among all the rounds $1 \le t \le T$. Thus, 0 is outputted a $1-r$ fraction of the rounds $1 \le t \le T$. It follows that the best fit normal to $\hat{\mu}_T$ has mean $r \cdot 1 + (1-r) \cdot 0 = r$ and variance $r \cdot 1 + (1-r) \cdot 0 - r^2 = r - r^2$. The resulting FID loss is

$$\left(r - \frac{1}{2}\right)^2 + \left(r - r^2 + \frac{1}{4} - \sqrt{r - r^2}\right) = r^2 - r + \frac{1}{4} + r - r^2 + \frac{1}{4} - \sqrt{r - r^2} = \frac{1}{2} - \sqrt{r - r^2}.$$

Now, suppose that $m' = m_1$. Then $x_t = 0$ for at most $T/4$ indices t, and hence $r \ge 3/4$. Similarly, if $m' = m_2$, then $x_t = 1$ for at most $T/4$ indices t, and hence $r \le 1/4$. Thus $r \notin \left(\frac{1}{4}, \frac{3}{4}\right)$. Therefore, the FID loss suffered by $\mathcal{A}$ is at least $\frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{16}} = \frac{2-\sqrt{3}}{4}$.

On the other hand, notice that in hindsight, m' will be such that $m'(p_t) = 1$ for exactly $T/2$ indices $t, 1 \le t \le T$. The corresponding FID loss to m' is $\frac{1}{2} - \sqrt{\frac{1}{2} - \frac{1}{4}} = \frac{1}{2} - \frac{1}{2} = 0$.

Therefore, the regret of $\mathcal{A}$ is at least $\frac{2-\sqrt{3}}{4}$. □

D EXTENSION TO COMPETING WITH THE BEST TOPIC SPLIT

D.1 PROBLEM SETTING

The setting is similar to the multi-armed bandit setting. The main two differences are that we now have the ability to condition our choice of model on the prompt, and there is now a set of experts which we can compare our algorithm against. Formally, we keep Assumption 2, Assumption 3, and Assumption 4 but discard Assumption 1, instead replacing it with the assumptions below.

We have K models, which we interpret as diffusion models. The learning process lasts for T days, with a prompt p_t coming in each day.

Assumption 5. *There exists a fixed prompt distribution p_D on prompt space $\mathcal{P}$ from which p_t are drawn iid.*

This should be regarded as the analog of the assumption of contexts being drawn iid in the contextual bandit setting. Note that this is still consistent with a multi-armed bandit setting because the learning algorithm will not depend directly on the values of p_t.

For each $i \in [K]$, model m_i maps an element of $\mathcal{P}$ to an element of $\mathcal{X}$, the output space. Note that in reality diffusion models would output in the image space but we will consider $\mathcal{X}$ to be the latent space, encoding the latent space transformation into m (Heusel et al., 2017b).

Assumption 6. *For each $i \in [K]$, $o \in [O]$, there exists a fixed model/topic distribution $D_{i,o}$ on $\mathcal{X}$, induced by m_i and p_D conditioned on being topic o.*

This should be regarded as the analog of the assumption of a stochastic loss distribution in the traditional multi-armed bandit setting.

We first try to study the setting in which the regret is based on topic splits, but the topic of each prompt is not revealed to the algorithm.

In particular, the goal is to design algorithms which achieve sub-linear regret, where the regret is

$$\min_{f \in \mathcal{F}} \left[\sum_{o \in [O]} \left[|X_{o,f}| \cdot L_{\text{total},o}(X_{o,f}) \right] - \sum_{o \in [O]} \min_{m \in [K]} \left[|X'_{o,f,m}| \cdot L_{\text{total},o}(X'_{o,f,m}) \right] \right]$$

where $X_{o,f} = \{x_t \text{ for } t \in [T] \text{ s.t. } f(p_t) = o\}$ and $X'_{o,f,m} = \{m_{m(f(p_t))}(p_t) \text{ for } t \in [T] \text{ s.t. } f(p_t) = o\}$ for all $o \in [O]$, $f \in \mathcal{F}$, and $m \in [K]$.

Lemma 4. *No algorithm achieves sub-linear regret for this problem.*

Proof. As discussed earlier, the FID loss satisfies our assumptions for the loss function so that suffices to disprove the more general statement. We construct a distribution in one dimension with only two experts, two models, and two topics that shows the negative result for FID loss.

Specifically, suppose that there are two prompts, p_1 and p_2. Suppose that model m_1 generates 1 in response to p_1 and 2 in response to p_2. Suppose that model m_2 generates 2 in response to p_1 and 1 in response to p_2.

Now, suppose that the prompt distribution p_D is such that $p(p_1) = p(p_2) = 1/2$. In addition, suppose that $\mathcal{N}_{\text{ref},o} = \mathcal{N}(o, 0)$ for $o \in [2]$.

The expert set is $\mathcal{F} = \{f_1, f_2\}$ where $f_1(p_1) = 1$, $f_1(p_2) = 2$, $f_2(p_1) = 2$, and $f_2(p_2) = 1$.

This can be interpreted as expert 1 believing that p_1 corresponds to topic 1 while expert 2 believes that p_1 corresponds to topic 2.

Consider the notion of agreeing with an expert, as playing the same model as that expert would suggest. Then, out of T rounds, we must disagree with either expert f_1 or expert f_2 in at least $T/2$ rounds, by the Generalized Pigeinhole Principle and the fact that the experts always output different values on the same prompt. The situation is symmetric in either case, so we will assume, without loss of generality, that we disagree with expert f_1 in at least $T/2$ rounds. Let d_1 be the number of disagreements with expert f_1 that occur on prompt p_1 (i.e. we play model m_2 while the expert suggests model m_1) and let d_2 be the number of disagreements with expert f_1 on model p_2.

Then, we have

$$\begin{aligned}\mathrm{Reg}_T &= \min_{f\in\mathcal{F}}\left[\sum_{o\in[O]}\left[|X_{o,f}|\cdot\mathrm{FID}(X_{o,f},\mathcal{N}_{\mathrm{ref},o})\right]-\sum_{o\in[O]}\min_{m\in[K]}\left[|X'_{o,f,m}|\cdot\mathrm{FID}(X'_{o,f,m},\mathcal{N}_{\mathrm{ref},o})\right]\right]\\ &\geq\left[\sum_{o\in[O]}\left[|X_{o,f_1}|\cdot\mathrm{FID}(X_{o,f_1},\mathcal{N}_{\mathrm{ref},o})\right]-\sum_{o\in[O]}\min_{m\in[K]}\left[|X'_{o,f_1,m}|\cdot\mathrm{FID}(X'_{o,f_1,m},\mathcal{N}_{\mathrm{ref},o})\right]\right]\\ &\geq\left[\sum_{o\in[O]}\left[|X_{o,f_1}|\cdot\mathrm{FID}(X_{o,f_1},\mathcal{N}_{\mathrm{ref},o})\right]\right]\end{aligned}$$

where the second term vanishes because for expert f_1, model m_1 achieves FID loss of 0 on prompt p_1 and model m_2 achieves FID loss of 0 on prompt p_2 (so these can be chosen in the respective minimum over models). Now, since the covariance part of FID is non-negative, we have that

$$\left[\sum_{o\in[O]}\left[|X_{o,f_1}|\cdot\mathrm{FID}(X_{o,f_1},\mathcal{N}_{\mathrm{ref},o})\right]\right]\geq\left[\sum_{o\in[O]}\left[|X_{o,f_1}|\cdot|\mu_{o,f_1}-o|_2^2\right]\right] \tag{38}$$

where μ_{o,f_1} is the mean generation that our algorithm produced when f_1 believed the prompt corresponded to topic o. Splitting the sum, we have

$$\left[\sum_{o\in[O]}\left[|X_{o,f_1}|\cdot|\mu_{o,f_1}-o|_2^2\right]\right]=|X_{1,f_1}|\cdot|\mu_{1,f_1}-1|_2^2+|X_{2,f_1}|\cdot|\mu_{2,f_1}-2|_2^2 \tag{39}$$

By our definitions of d_1 and d_2, we have

$$|X_{1,f_1}|\cdot|\mu_{1,f_1}-1|_2^2+|X_{2,f_1}|\cdot|\mu_{2,f_1}-2|_2^2\geq|X_{1,f_1}|\cdot\left(\frac{d_1}{T}\right)^2+|X_{2,f_1}|\cdot\left(\frac{d_2}{T}\right)^2 \tag{40}$$

Since $d_1+d_2\geq\frac{T}{2}$, by the Pigeonhole Principle, at least one of them must be at least $\frac{T}{4}$. With probability at least $\frac{1}{2}$, both $|X_{1,f_1}|\geq\frac{T}{2}-\sqrt{T\log(8)}$ and $|X_{2,f_1}|\geq\frac{T}{2}-\sqrt{T\log(8)}$. Then,

$$\mathbb{E}\left[|X_{1,f_1}|\cdot\left(\frac{d_1}{T}\right)^2+|X_{2,f_1}|\cdot\left(\frac{d_2}{T}\right)^2\right]\geq\left[\frac{T}{2}-\sqrt{T\log(8)}\right]\cdot\frac{T^2}{16T^2}$$

so for any $T>16\log(8)\approx33$, the expected regret is at least $\frac{T}{4}\cdot\frac{1}{16}=\frac{T}{64}$, as desired. □

This shows that the algorithm needs topic feedback to do well, in a precise sense. Specifically, we study a similar notion of regret as before but where the true topic o_t is provided at the end of each round.

We now state the formal learning procedure.

Algorithm 4 Learning Procedure

1: Receive loss functions $L_{\mathrm{total},o}$ for all $o\in[O]$
2: Receive expert set $\mathcal{F}$
3: **for** $t=1$ to T **do**
4: Learner observes prompt p_t
5: Learner picks $i_t\in[K]$
6: Learner observes image generation $x_t=m_{i_t}(p_t)$
7: Learner observes topic o_t
8: Learner suffers loss $\sum_{o\in[O]}|X_o|\cdot L_{\mathrm{total},o}(X_o)$ where $X_o=\{x_t \text{ for } t\in[T] \text{ s.t. } o_t=o\}$ for all $o\in[O]$.

The goal is to design algorithms which achieve sub-linear regret, where the normalized regret is

$$\mathrm{Reg}_T := \frac{1}{T}\sum_{o\in[O]} |X_o| \cdot L_{\mathrm{total},o}(X_o) - \min_{f\in\mathcal{F}} \min_{\pi\in[K]^O} \frac{1}{T}\sum_{o\in[O]} |X_o| \cdot L_{\mathrm{total},o}(X'_{o,f,\pi}) \tag{41}$$

where $X'_{o,f,\pi} = \{m_{\pi(f(p_t))}(p_t)$ for $t \in [T]$ s.t. $o_t = o\}$ for all $o \in [O]$, $f \in \mathcal{F}$, and $\pi \in [K]^O$. Consider the following algorithm.

Algorithm 5 Explore then Commit (Topic Version)

Set $C = \lfloor T^{\frac{p}{p+1}} \rfloor$, where $p \geq 1$ is from Assumption 2
for each topic $o \in [O]$ **do**
  for each expert $f \in \mathcal{F}$ and mapping $\pi \in [K]^O$ **do**
    Create generation set $S_{f,\pi,o} = \varnothing$
  Create topic count $c_o = 0$
for $t = 1$ to C **do**
  Observe prompt p_t
  Play random model $i_t \sim \mathrm{Unif}([K])$
  Observe generation $x_t = m_{i_t}(p_t)$
  Observe topic o_t
  for each expert $f \in \mathcal{F}$ and mapping $\pi \in [K]^O$ **do**
    if $\pi(f(p_t)) = i_t$ **then**
      Add x_t to S_{f,π,o_t}
  Increment c_{o_t}
Let $(f^*, \pi^*) = \mathrm{argmin}_{f\in\mathcal{F},\pi\in[K]^O} \sum_{o\in[O]} [c_o \cdot L_{\mathrm{total},o}(S_{f,\pi,o})]$
for $t = C+1$ to T **do**
  Observe prompt p_t
  Set $i_t = \pi^*(f^*(p_t))$
  Observe generation $x_t = m_{i_t}(p_t)$

Theorem 7. *Suppose Assumption 2, Assumption 3, and Assumption 4 hold. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 5 achieves the following regret bound:*

$$\mathrm{Reg}_T \leq B_{\mathrm{total}} T^{-\frac{1}{p+1}} + 2B_{\mathrm{total}} O(2K+1)\sqrt{\frac{L}{2}} T^{-\frac{p}{2(p+1)}} + 2C_{\mathrm{total}} K (OL)^{1/p} T^{-\frac{1}{p+1}}, \tag{42}$$

where $L = \log\left(\frac{12|\mathcal{F}|K^O O}{\delta}\right)$. In particular, for any $p \geq 2$, with probability at least $1-\delta$,

$$\mathrm{Reg}_T \leq O\left(\left[B_{\mathrm{total}} OK\sqrt{L} + C_{\mathrm{total}} K (OL)^{1/p}\right] T^{-\frac{1}{p+1}}\right). \tag{43}$$

When $p = 2$ (corresponding to sub-Gaussian concentration, such as for FID), with probability at least $1-\delta$,

$$\begin{aligned}\mathrm{Reg}_T &\leq \frac{1}{T^{1/3}}\left[B_{\mathrm{total}} + 2\sqrt{L}\left(\sqrt{2}B_{\mathrm{total}} O(2K+1) + C_{\mathrm{total}} K\sqrt{O}\right)\right] \\ &= O\left(\left[B_{\mathrm{total}} OK + C_{\mathrm{total}} K\sqrt{O}\right]\sqrt{L}\, T^{-1/3}\right).\end{aligned} \tag{44}$$

Proof. For any fixed expert $f \in \mathcal{F}$, mapping $\pi : [O] \to [K]$, and topic $o \in [O]$, let $S_{f,\pi,o}$ be the set of generations collected during the C exploration rounds where $o_t = o$ and the randomly played model matches the mapping's suggestion, $\pi(f(p_t)) = i_t$. Let c_o denote the number of rounds during exploration where $o_t = o$. Since the prompts p_t are drawn i.i.d., the true topics o_t follow a marginal distribution $p_o = \Pr(o_t = o)$. The expected value of c_o is Cp_o. Furthermore, since the models $i_t \sim \mathrm{Unif}([K])$ are chosen uniformly at random and independently, the expected size of $S_{f,\pi,o}$ is:

$$\mathbb{E}[|S_{f,\pi,o}|] = \sum_{t=1}^{C} \Pr(o_t = o)\Pr(\pi(f(p_t)) = i_t) = \frac{C}{K} p_o.$$

Both c_o and $|S_{f,\pi,o}|$ are sums of C independent indicator variables. By Hoeffding's inequality, their deviations from their expectations are bounded. Applying a union bound over all $o \in [O]$ for c_o, and over all $f \in \mathcal{F}, \pi : [O] \to [K], o \in [O]$ for $|S_{f,\pi,o}|$, we have that with probability at least $1 - \delta/3$:

$$|c_o - Cp_o| \le \sqrt{\frac{C}{2}L} \quad \text{and} \quad \left||S_{f,\pi,o}| - \frac{C}{K}p_o\right| \le \sqrt{\frac{C}{2}L},$$

where $L = \log\left(\frac{12|\mathcal{F}|K^O O}{\delta}\right)$.

Next, we bound the error in estimating the per-round expected true total loss $\mathcal{L}(f,\pi) = \sum_{o\in[O]} p_o L_{\text{total},o}(\mathcal{D}_{f,\pi,o})$, where $\mathcal{D}_{f,\pi,o}$ is the true distribution induced by mapping π and expert f on topic o. The algorithm selects the pair (f^*, π^*) that minimizes the empirical estimator:

$$\hat{\mathcal{L}}(f,\pi) = \frac{1}{C}\sum_{o\in[O]} c_o L_{\text{total},o}(S_{f,\pi,o}).$$

We relate this estimator to a sample-weighted estimator $\hat{\mathcal{L}}_S(f,\pi) = \frac{K}{C}\sum_{o\in[O]} |S_{f,\pi,o}| L_{\text{total},o}(S_{f,\pi,o})$. The difference is bounded by the concentration of c_o and $|S_{f,\pi,o}|$ around their shared parameter p_o. Using the triangle inequality and the bounds on c_o and $|S_{f,\pi,o}|$, we have:

$$|c_o - K|S_{f,\pi,o}|| \le |c_o - Cp_o| + K\left||S_{f,\pi,o}| - \frac{C}{K}p_o\right| \le (K+1)\sqrt{\frac{C}{2}L}.$$

Using Assumption 4 ($L_{\text{total},o} \le B_{\text{total}}$), the difference between the estimators is:

$$\begin{aligned}|\hat{\mathcal{L}}(f,\pi) - \hat{\mathcal{L}}_S(f,\pi)| &\le \frac{1}{C}\sum_{o\in[O]} |c_o - K|S_{f,\pi,o}||L_{\text{total},o}(S_{f,\pi,o}) \\ &\le \frac{1}{C}\sum_{o\in[O]} (K+1)\sqrt{\frac{C}{2}L}\cdot B_{\text{total}} = B_{\text{total}} O(K+1)\sqrt{\frac{L}{2C}}.\end{aligned}$$

Now, we bound the estimation error $|\hat{\mathcal{L}}_S(f,\pi) - \mathcal{L}(f,\pi)|$, splitting it into sample size deviation and empirical total loss concentration:

$$\begin{aligned}|\hat{\mathcal{L}}_S(f,\pi) - \mathcal{L}(f,\pi)| \le &\sum_{o\in[O]}\left|\frac{K}{C}|S_{f,\pi,o}| - p_o\right| L_{\text{total},o}(\mathcal{D}_{f,\pi,o}) \\ &+ \frac{K}{C}\sum_{o\in[O]} |S_{f,\pi,o}||L_{\text{total},o}(S_{f,\pi,o}) - L_{\text{total},o}(\mathcal{D}_{f,\pi,o})|.\end{aligned}$$

For the first term, applying the Hoeffding bound and $L_{\text{total},o} \le B_{\text{total}}$, we get:

$$\sum_{o\in[O]}\left|\frac{K}{C}|S_{f,\pi,o}| - p_o\right| L_{\text{total},o}(\mathcal{D}_{f,\pi,o}) \le B_{\text{total}}\sum_{o\in[O]}\frac{K}{C}\sqrt{\frac{CL}{2}} = B_{\text{total}} OK\sqrt{\frac{L}{2C}}.$$

For the second term, we apply Assumption 2. For each fixed $f \in \mathcal{F}$, $\pi \in [K]^O$, and $o \in [O]$, the set $S_{f,\pi,o}$ comprises $|S_{f,\pi,o}|$ i.i.d. generations drawn from $\mathcal{D}_{f,\pi,o}$. By Assumption 2 with failure probability $\frac{\delta}{3|\mathcal{F}|K^O O}$, taking a union bound over all $f \in \mathcal{F}$, $\pi \in [K]^O$, and $o \in [O]$, with probability at least $1 - \delta/3$,

$$|L_{\text{total},o}(S_{f,\pi,o}) - L_{\text{total},o}(\mathcal{D}_{f,\pi,o})| \le C_{\text{total}}\left(\frac{L}{|S_{f,\pi,o}|}\right)^{1/p},$$

where $L \ge \log\left(\frac{3|\mathcal{F}|K^O O}{\delta}\right)$. Multiplying by $|S_{f,\pi,o}|$ and summing across all topics $o \in [O]$, we obtain:

$$\frac{K}{C}\sum_{o\in[O]} |S_{f,\pi,o}||L_{\text{total},o}(S_{f,\pi,o}) - L_{\text{total},o}(\mathcal{D}_{f,\pi,o})| \le C_{\text{total}} L^{1/p}\frac{K}{C}\sum_{o\in[O]} |S_{f,\pi,o}|^{1-1/p}.$$

To bound this sum across the topics, we apply Hölder's inequality. Intuitively, if each individual topic's loss were estimated to within an accuracy of ε, summing across O topics would contribute an

aggregate error of $\varepsilon \cdot O$. However, here the sample size $|S_{f,\pi,o}| \approx \frac{C}{K} p_o$ depends on each topic's specific probability p_o. Hölder's inequality enables bounding the aggregate error without needing lower bounds on individual topic probabilities. Specifically, for $p = 1$, $1 - 1/p = 0$, so $\sum_{o\in[O]} |S_{f,\pi,o}|^0 = O = O^{1/1}$. For $p > 1$, let $q = \frac{p}{p-1}$ be the Hölder conjugate of p so that $\frac{1}{p} + \frac{1}{q} = 1$. Applying Hölder's inequality with conjugate exponents p and q to the vectors $(1)_{o\in[O]}$ and $(|S_{f,\pi,o}|^{1-1/p})_{o\in[O]}$, we have:

$$\sum_{o\in[O]} 1 \cdot |S_{f,\pi,o}|^{1-1/p} \le \left(\sum_{o\in[O]} 1^p \right)^{1/p} \left(\sum_{o\in[O]} \left(|S_{f,\pi,o}|^{\frac{p-1}{p}} \right)^{\frac{p}{p-1}} \right)^{\frac{p-1}{p}}$$
$$= O^{1/p} \left(\sum_{o\in[O]} |S_{f,\pi,o}| \right)^{1-1/p}.$$

Since each exploration round generates exactly one sample belonging to a single topic o_t, the sets $\{S_{f,\pi,o}\}_{o\in[O]}$ are disjoint subsets of the C exploration samples, so $\sum_{o\in[O]} |S_{f,\pi,o}| \le C$. Therefore,

$$\sum_{o\in[O]} |S_{f,\pi,o}|^{1-1/p} \le O^{1/p} C^{1-1/p}.$$

Multiplying by $\frac{K}{C} C_{\text{total}} L^{1/p}$, we obtain:

$$\frac{K}{C} \sum_{o\in[O]} |S_{f,\pi,o}| |L_{\text{total},o}(S_{f,\pi,o}) - L_{\text{total},o}(\mathcal{D}_{f,\pi,o})| \le \frac{K}{C} C_{\text{total}} L^{1/p} \cdot O^{1/p} C^{1-1/p} = C_{\text{total}} K \left(\frac{OL}{C} \right)^{1/p}.$$

Summing the estimator deviation and concentration bounds, with probability at least $1 - 2\delta/3$, the total estimation error $|\hat{\mathcal{L}}(f,\pi) - \mathcal{L}(f,\pi)|$ for every pair (f,π) is at most:

$$\varepsilon_{\text{est}} := B_{\text{total}} O(2K+1) \sqrt{\frac{L}{2C}} + C_{\text{total}} K \left(\frac{OL}{C} \right)^{1/p}. \tag{45}$$

Finally, we bound the overall regret by separating the exploration and exploitation phases using convexity (Assumption 3). For each topic $o \in [O]$, the set of generations X_o produced by the algorithm over all T rounds is partitioned into exploration samples X_o^{explore} (of size c_o) and exploitation samples X_o^{exploit} (of size $|X_o| - c_o$). By Assumption 3, the empirical distribution of X_o is a convex combination of the exploration and exploitation empirical distributions, giving:

$$L_{\text{total},o}(X_o) \le \frac{c_o}{|X_o|} L_{\text{total},o}(X_o^{\text{explore}}) + \frac{|X_o| - c_o}{|X_o|} L_{\text{total},o}(X_o^{\text{exploit}}).$$

Multiplying by $\frac{|X_o|}{T}$ and summing across all topics $o \in [O]$, we decompose the learner's total loss:

$$\frac{1}{T} \sum_{o\in[O]} |X_o| L_{\text{total},o}(X_o) \le \frac{1}{T} \sum_{o\in[O]} c_o L_{\text{total},o}(X_o^{\text{explore}}) + \frac{1}{T} \sum_{o\in[O]} (|X_o| - c_o) L_{\text{total},o}(X_o^{\text{exploit}}).$$

We bound the two terms as follows:

1. **Exploration phase:** By Assumption 4, $L_{\text{total},o}(X_o^{\text{explore}}) \le B_{\text{total}}$. Since $\sum_{o\in[O]} c_o = C$, the exploration phase contributes at most:

$$\frac{1}{T} \sum_{o\in[O]} c_o L_{\text{total},o}(X_o^{\text{explore}}) \le \frac{B_{\text{total}}}{T} \sum_{o\in[O]} c_o = B_{\text{total}} \frac{C}{T} \le B_{\text{total}} T^{-\frac{1}{p+1}}.$$

   Due to the small number of exploration rounds $C = \lfloor T^{\frac{p}{p+1}} \rfloor \ll T$ and bounded loss, exploration has minimal effect on regret.

2. **Exploitation phase:** During the $T - C$ exploitation rounds, the algorithm plays the empirical best expert and mapping pair (f^*, π^*). Let $(f_{\text{opt}}, \pi_{\text{opt}}) = \text{argmin}_{f\in\mathcal{F}, \pi\in[K]^O} \mathcal{L}(f,\pi)$ denote

the population optimal pair. Since (f^*, π^*) minimizes $\hat{\mathcal{L}}$ and $|\hat{\mathcal{L}}(f,\pi) - \mathcal{L}(f,\pi)| \le \varepsilon_{\text{est}}$ uniformly:

$$\mathcal{L}(f^*,\pi^*) \le \hat{\mathcal{L}}(f^*,\pi^*) + \varepsilon_{\text{est}} \le \hat{\mathcal{L}}(f_{\text{opt}},\pi_{\text{opt}}) + \varepsilon_{\text{est}} \le \mathcal{L}(f_{\text{opt}},\pi_{\text{opt}}) + 2\varepsilon_{\text{est}}.$$

By convexity over the topics, the loss during the exploitation phase is bounded by the population loss plus the estimation error, yielding:

$$\frac{1}{T}\sum_{o\in[O]}(|X_o| - c_o)L_{\text{total},o}(X_o^{\text{exploit}}) - \min_{f\in\mathcal{F},\pi\in[K]^O}\frac{1}{T}\sum_{o\in[O]}|X_o|L_{\text{total},o}(X'_{o,f,\pi}) \le 2\varepsilon_{\text{est}}.$$

Combining the exploration and exploitation bounds, with probability at least $1-\delta$, the cumulative normalized regret satisfies:

$$\begin{aligned}\text{Reg}_T &\le B_{\text{total}}\frac{C}{T} + 2\varepsilon_{\text{est}} \\ &\le B_{\text{total}}T^{-\frac{1}{p+1}} + 2B_{\text{total}}O(2K+1)\sqrt{\frac{L}{2C}} + 2C_{\text{total}}K\left(\frac{OL}{C}\right)^{1/p}.\end{aligned}$$

Substituting $C = \lfloor T^{\frac{p}{p+1}} \rfloor \le T^{\frac{p}{p+1}}$ into ε_{est}:

$$\text{Reg}_T \le B_{\text{total}}T^{-\frac{1}{p+1}} + 2B_{\text{total}}O(2K+1)\sqrt{\frac{L}{2}}T^{-\frac{p}{2(p+1)}} + 2C_{\text{total}}K(OL)^{1/p}T^{-\frac{1}{p+1}}.$$

For $p \ge 2$, we have $-\frac{p}{2(p+1)} \le -\frac{1}{p+1}$, so the dominant rate is $T^{-\frac{1}{p+1}}$, giving:

$$\text{Reg}_T \le O\left(\left[B_{\text{total}}OK\sqrt{L} + C_{\text{total}}K(OL)^{1/p}\right]T^{-\frac{1}{p+1}}\right).$$

In particular, for $p = 2$ (corresponding to sub-Gaussian concentration, such as FID), both exponents $-\frac{p}{2(p+1)}$ and $-\frac{1}{p+1}$ equal $-1/3$, yielding:

$$\begin{aligned}\text{Reg}_T &\le \frac{1}{T^{1/3}}\left[B_{\text{total}} + 2\sqrt{L}\left(\sqrt{2}B_{\text{total}}O(2K+1) + C_{\text{total}}K\sqrt{O}\right)\right] \\ &= O\left(\left[B_{\text{total}}OK + C_{\text{total}}K\sqrt{O}\right]\sqrt{L}T^{-1/3}\right),\end{aligned}$$

completing the proof. □

Consider the following notion of an ERM oracle for this problem:

$$O_{ERM}(\{(p_t,i_t,x_t,o_t)\}_{t=1}^T) = \operatorname*{argmin}_{f\in\mathcal{F},\pi\in[K]^O}\sum_{o\in[O]}\left[c_o \cdot L_{\text{total},o}(S_{f,\pi,o})\right].$$

where

$$S_{f,\pi,o} = \{x_t : t\in[T], \pi(f(p_t)) = i_t\} \text{ and } c_o = |\{t\in[T] : o_t = o\}|.$$

Lemma 5. *Algorithm 5 is an ERM-oracle efficient algorithm.*

Proof. This can be seen from the fact that when Algorithm 5 has access to an ERM oracle, Line 11 can be directly implemented by calling that oracle on the true values of p_t, i_t, x_t and o_t. The other lines involving $S_{f,\pi,o}$ can be removed. □

Theorem 8. *Suppose Assumption 2, Assumption 3, and Assumption 4 hold. Then, Algorithm 6 achieves the following bound on expected regret:*

$$\begin{aligned}\mathbb{E}[\text{Reg}_T] \le{}& \frac{B_{\text{total}}\cdot\left(36K^2L_T^2+1\right)}{T} + \frac{2\sqrt{2}B_{\text{total}}O(2K+1)\sqrt{L_T}}{\sqrt{2}-1}\frac{1}{\sqrt{T}} \\ &+ \frac{2^{2+1/p}C_{\text{total}}K(OL_T)^{1/p}}{2^{1-1/p}-1}\frac{1}{T^{1/p}},\end{aligned} \tag{46}$$

Algorithm 6 Improving dependence on T (Topic Version)

Initialize history $H_0 = \varnothing$
Initialize $m = 0$
for $t = 1$ to T **do**
  Observe prompt p_t
  if $t = 2^m$ **then**
    Increment m
    Obtain $(\hat{f}_m, \hat{\pi}_m) = O_{ERM}(H_{t-1})$ (for $m = 1$, choose any $f \in \mathcal{F}, \pi \in [K]^O$ arbitrarily)
  Sample a random model $i_t^{(1)}$ uniformly from $[K]$
  Observe generation $x_t^{(1)} = m_{i_t^{(1)}}(p_t)$
  Select model $i_t^{(2)} = \hat{\pi}_m(\hat{f}_m(p_t))$
  Observe generation $x_t^{(2)} = m_{i_t^{(2)}}(p_t)$
  Observe topic o_t
  Update history $H_t = H_{t-1} \cup \{(p_t, i_t^{(1)}, x_t^{(1)}, o_t)\}$

where $L_T = \log(12|\mathcal{F}|K^O OT)$. *In particular, for any* $p \geq 2$,

$$\mathbb{E}[\mathrm{Reg}_T] \leq O\left(\left[B_{\mathrm{total}} OK\sqrt{L_T} + \frac{C_{\mathrm{total}} K (OL_T)^{1/p}}{2^{1-1/p} - 1}\right] T^{-1/p}\right). \tag{47}$$

When $p = 2$ *(corresponding to sub-Gaussian concentration, such as for FID), this yields:*

$$\mathbb{E}[\mathrm{Reg}_T] \leq O\left(\left[B_{\mathrm{total}} OK + C_{\mathrm{total}} K\sqrt{O}\right]\sqrt{\frac{L_T}{T}}\right). \tag{48}$$

Proof. We assume without loss of generality that $T = 2^M - 1$ for some integer $M = \log_2(T+1)$; if T is not of this form, the final epoch is simply truncated, which changes the bound by at most a factor of 2. The T rounds are partitioned into M epochs, where each epoch $m \in [M]$ consists of rounds $t \in [2^{m-1}, 2^m - 1]$ and has length $T_m = 2^{m-1}$. At the beginning of epoch m (step $t = 2^{m-1}$), the history H_{t-1} contains $C_m = 2^{m-1} - 1$ exploration samples collected by playing models chosen uniformly at random from $[K]$. The algorithm computes $(\hat{f}_m, \hat{\pi}_m) = O_{ERM}(H_{2^{m-1}-1})$ and commits to this pair throughout epoch m.

First, we decompose the total regret across epochs and topics using Assumption 3. For each topic $o \in [O]$ and epoch $m \in [M]$, let $X_{o,m} = \{x_t^{(2)} : 2^{m-1} \leq t \leq 2^m - 1, o_t = o\}$ be the user-facing generations in epoch m on topic o, and let $n_{o,m} = |X_{o,m}|$, so that $\sum_{o\in[O]} n_{o,m} = 2^{m-1}$ and $\sum_{m=1}^M n_{o,m} = |X_o|$. For each topic o, the total generation set is $X_o = \bigcup_{m=1}^M X_{o,m}$. By Assumption 3, the empirical distribution of X_o is a convex combination of the epoch distributions:

$$L_{\mathrm{total},o}(X_o) \leq \sum_{m=1}^M \frac{n_{o,m}}{|X_o|} L_{\mathrm{total},o}(X_{o,m}).$$

Multiplying by $\frac{|X_o|}{T}$ and summing over all topics $o \in [O]$, we obtain:

$$\frac{1}{T}\sum_{o\in[O]} |X_o| L_{\mathrm{total},o}(X_o) \leq \sum_{m=1}^M \frac{2^{m-1}}{T}\left(\sum_{o\in[O]} \frac{n_{o,m}}{2^{m-1}} L_{\mathrm{total},o}(X_{o,m})\right).$$

Let $(f^\star, \pi^\star) = \mathrm{argmin}_{f\in\mathcal{F},\pi\in[K]^O} \frac{1}{T}\sum_{o\in[O]} |X_o| L_{\mathrm{total},o}(X'_{o,f,\pi})$ denote the optimal pair in hindsight, and let $(f_{\mathrm{opt}}, \pi_{\mathrm{opt}}) = \mathrm{argmin}\, \mathcal{L}(f,\pi)$ denote the population optimal pair, where $\mathcal{L}(f,\pi) = \sum_{o\in[O]} p_o L_{\mathrm{total},o}(\mathcal{D}_{f,\pi,o})$. Since $\sum_{m=1}^M \frac{2^{m-1}}{T} = 1$, the cumulative normalized regret decomposes as:

$$\mathrm{Reg}_T \leq \sum_{m=1}^M \frac{2^{m-1}}{T}\left(\sum_{o\in[O]} \frac{n_{o,m}}{2^{m-1}} L_{\mathrm{total},o}(X_{o,m}) - \frac{1}{T}\sum_{o\in[O]} |X_o| L_{\mathrm{total},o}(X'_{o,f^\star,\pi^\star})\right).$$

Next, we bound the estimation error in epoch m. In history $H_{2^{m-1}-1}$, each prompt $p_\tau \sim p_D$ is i.i.d., each topic $o_\tau \sim (p_o)_{o\in[O]}$, and each exploration model $i_\tau^{(1)} \sim \mathrm{Unif}([K])$ is drawn independently. For each $f \in \mathcal{F}, \pi \in [K]^O, o \in [O]$, let $S^{(m)}_{f,\pi,o}$ be the subset of $H_{2^{m-1}-1}$ where topic was o and $i_\tau^{(1)} = \pi(f(p_\tau))$, and let $c_o^{(m)}$ be the count of topic o. By Hoeffding's inequality and Assumption 2, applying Hölder's inequality across the topic splits exactly as in the proof of Theorem 7, with $L_T = \log(12|\mathcal{F}|K^O OT)$, with probability at least $1-\delta_T$ (where $\delta_T = 1/T$), the estimation error uniformly satisfies:

$$\left|\hat{\mathcal{L}}_m(f,\pi) - \mathcal{L}(f,\pi)\right| \le \varepsilon_{\text{est}}(m) := B_{\text{total}} O(2K+1)\sqrt{\frac{L_T}{2C_m}} + C_{\text{total}} K \left(\frac{OL_T}{C_m}\right)^{1/p}$$

for all $f \in \mathcal{F}, \pi \in [K]^O$, where $C_m = 2^{m-1}-1$. Because $(\hat{f}_m, \hat{\pi}_m) = \operatorname{argmin} \hat{\mathcal{L}}_m$, we have $\mathcal{L}(\hat{f}_m, \hat{\pi}_m) \le \mathcal{L}(f_{\text{opt}}, \pi_{\text{opt}}) + 2\varepsilon_{\text{est}}(m)$. During epoch m, $(\hat{f}_m, \hat{\pi}_m)$ is fixed, so $X_{o,m}$ contains $n_{o,m}$ i.i.d. samples from $\mathcal{D}_{\hat{f}_m,\hat{\pi}_m,o}$. Taking into account concentration over the epoch rounds and the T hindsight rounds, the excess loss in epoch m is bounded by $2\varepsilon_{\text{est}}(m)$.

Now, we split the epochs into initial epochs with small sample sizes and subsequent epochs: Let $\tau_0 = 18K^2L_T^2$. For initial epochs where $C_m \le \tau_0$, we use the trivial upper bound from Assumption 4, noting that the loss difference is at most B_{total}. The total number of rounds in these initial epochs is $\sum_{m:C_m\le\tau_0} 2^{m-1} \le 2\tau_0$, contributing at most:

$$\sum_{m:C_m\le\tau_0} \frac{2^{m-1}}{T} B_{\text{total}} \le \frac{2B_{\text{total}}\tau_0}{T} = \frac{B_{\text{total}} \cdot 36K^2L_T^2}{T}.$$

For epochs where $C_m > \tau_0$, we have $C_m = 2^{m-1}-1 \ge 2^{m-2}$. We sum $2\varepsilon_{\text{est}}(m)$ weighted by $\frac{2^{m-1}}{T}$:

$$\sum_{m:C_m>\tau_0} \frac{2^{m-1}}{T} \cdot 2\varepsilon_{\text{est}}(m) \le \frac{2B_{\text{total}}O(2K+1)\sqrt{L_T}}{\sqrt{2}T} \sum_{m=1}^{M} \frac{2^{m-1}}{\sqrt{C_m}} + \frac{2C_{\text{total}}K(OL_T)^{1/p}}{T} \sum_{m=1}^{M} \frac{2^{m-1}}{C_m^{1/p}}.$$

Using $C_m \ge 2^{m-2}$, we bound the two summations using geometric series:

1. For the first summation, since $\frac{2^{m-1}}{\sqrt{C_m}} \le \sqrt{2}\cdot 2^{(m-1)/2}$:

$$\sum_{m=1}^{M} 2^{(m-1)/2} = \frac{2^{M/2}-1}{\sqrt{2}-1} \le \frac{\sqrt{2T}}{\sqrt{2}-1}.$$

   Multiplying by $\frac{2B_{\text{total}}O(2K+1)\sqrt{L_T}}{\sqrt{2}T} \cdot \sqrt{2}$ yields:

$$\frac{2\sqrt{2}B_{\text{total}}O(2K+1)\sqrt{L_T}}{\sqrt{2}-1} \frac{1}{\sqrt{T}}.$$

2. For the second summation, since $\frac{2^{m-1}}{C_m^{1/p}} \le 2^{1/p}\cdot 2^{(m-1)(1-1/p)}$:

$$\sum_{m=1}^{M} 2^{(m-1)(1-1/p)} = \frac{2^{M(1-1/p)}-1}{2^{1-1/p}-1} \le \frac{(2T)^{1-1/p}}{2^{1-1/p}-1}.$$

   Multiplying by $\frac{2C_{\text{total}}K(OL_T)^{1/p}}{T} \cdot 2^{1/p}$ yields:

$$\frac{2^{2+1/p}C_{\text{total}}K(OL_T)^{1/p}}{2^{1-1/p}-1} \frac{1}{T^{1/p}}.$$

Under the failure event of probability at most $\delta_T = 1/T$, the regret is bounded by B_{total}, contributing at most $\frac{B_{\text{total}}}{T}$ to the expectation. Combining all terms, the expected regret is bounded by:

$$\mathbb{E}[\mathrm{Reg}_T] \le \frac{B_{\text{total}} \cdot \left(36K^2L_T^2+1\right)}{T} + \frac{2\sqrt{2}B_{\text{total}}O(2K+1)\sqrt{L_T}}{\sqrt{2}-1} \frac{1}{\sqrt{T}}$$

$$+\frac{2^{2+1/p}C_{\text{total}}K(OL_T)^{1/p}}{2^{1-1/p}-1}\frac{1}{T^{1/p}}.$$

For $p \geq 2$, the term $T^{-1/p}$ dominates $T^{-1/2}$, yielding the rate $\tilde{O}(T^{-1/p})$. In particular, when $p = 2$, both terms scale as $\frac{1}{\sqrt{T}}$, giving:

$$\mathbb{E}[\text{Reg}_T] \leq O\left(\left[B_{\text{total}}OK + C_{\text{total}}K\sqrt{O}\right]\sqrt{\frac{L_T}{T}}\right),$$

which completes the proof. □

E Experiment Details

In this section, we provide implementation details for the synthetic contextual distribution-matching experiment presented in Section 5 of the main text.

Data Generation

We simulate an online model routing problem over $T = 2500$ rounds with $O = 4$ discrete topics and $K = 4$ candidate generative models. The output representation space has dimension $d = 8$, representing latent image embeddings. At each round $t \in [T]$, a topic $o_t \in \{0, \ldots, O-1\}$ is drawn uniformly at random, and the context vector is its one-hot encoding $z_t = \mathbf{e}_{o_t} \in \mathbb{R}^O$. The target distribution is an isotropic Gaussian $\mathcal{N}(\mu_{\text{ref}}, \Sigma_{\text{ref}})$ with mean $\mu_{\text{ref}} = \mathbf{0} \in \mathbb{R}^d$ and covariance $\Sigma_{\text{ref}} = 0.2I_d$. Given topic o_t, each candidate model $k \in \{0, \ldots, K-1\}$ generates an output $X_{t,k} \in \mathbb{R}^d$ from a base stochastic noise $\xi_t \sim \mathcal{N}(\mathbf{0}, 0.2I_d)$:

$$X_{t,k} = \begin{cases} \xi_t, & \text{if } k = o_t \quad \text{(topic expert)}, \\ 1.5\xi_t + 2 \cdot \mathbf{e}_{(k \bmod d)}, & \text{if } k \neq o_t. \end{cases}$$

The specialized expert matches the target distribution perfectly ($\mathbb{E}[X_{t,k}] = \mu_{\text{ref}}$, $\mathbb{V}(X_{t,k}) = \Sigma_{\text{ref}}$), so an optimal routing policy achieves population FID equal to 0.

Algorithm Implementation

The routing policy is parameterized by weights $W \in \mathbb{R}^{K \times O}$ initialized to zero, outputting selection probabilities via a softmax distribution with temperature $\tau = 0.3$: $q_k(z_t; W) = \frac{\exp((Wz_t)_k/\tau)}{\sum_{j=1}^K \exp((Wz_t)_j/\tau)}$. Policy parameters W are updated every batch of $B = 25$ rounds by performing 30 SGD steps using the AdamW optimizer (learning rate $\eta = 0.03$, weight decay 10^{-4}, mini-batch size 128) on a differentiable weighted FID loss:

$$\mathcal{L}(W) = \|\hat{\mu}(W) - \mu_{\text{ref}}\|_2^2 + \text{tr}(\hat{\Sigma}_\varepsilon(W)) + \text{tr}(\Sigma_{\text{ref},\varepsilon}) - 2\text{tr}\left(\left(\Sigma_{\text{ref},\varepsilon}^{1/2}\hat{\Sigma}_\varepsilon(W)\Sigma_{\text{ref},\varepsilon}^{1/2}\right)^{1/2}\right) + 10^{-4}\|W\|_F^2.$$

Here, $\hat{\mu}(W)$ and $\hat{\Sigma}(W)$ are importance-weighted empirical estimates computed from the historical observation buffer H using weights $a_\tau(W) = q_{i_\tau}(z_\tau; W)$, regularized via $\hat{\Sigma}_\varepsilon(W) = \frac{1}{2}(\hat{\Sigma}(W) + \hat{\Sigma}(W)^\top) + \varepsilon_{\text{stab}}I_d$ and $\Sigma_{\text{ref},\varepsilon} = \Sigma_{\text{ref}} + \varepsilon_{\text{stab}}I_d$ with $\varepsilon_{\text{stab}} = 10^{-5}$.

Learning Format Here we further describe the algorithms compared in Figure 1.

- **Algorithm 1:** At each round t, an auxiliary exploratory model $i'_t \sim \text{Uniform}([K])$ is queried (not shown to the user) to record an unbiased transition $(z_t, m_{i'_t}(p_t), i'_t)$ into history buffer H_t. The policy W is trained exclusively on this unbiased buffer. The user-facing generation is sampled from the current policy: $i_t \sim q(z_t; W)$.
- **Algorithm 2:** At each round t, the algorithm selects only one model $i_t \sim q(z_t; W)$, displays it to the user, and stores $(z_t, m_{i_t}(p_t), i_t)$ into H_t. The policy W is trained on this same data.

Performance is assessed by the empirical FID of the running sequence of deployed generations $(x_1, \ldots, x_t)$ displayed to the user against $(\mu_{\text{ref}}, \Sigma_{\text{ref}})$.

F Supporting Lemmas on Convexity and Concentration

Lemma 6. *Consider any distribution μ on $\mathbb{R}^d$ bounded in $\mathbb{B}(0,R)$. Letting $\hat{\mu}_n$ denote the product distribution formed by n i.i.d. draws from μ, we have with probability at least $1-\delta$,*

$$|\mathrm{FID}(\hat{\mu}_n, \mu_{\mathrm{ref}}) - \mathrm{FID}(\mu, \mu_{\mathrm{ref}})| \le C_1(d,R)\sqrt{\frac{\log(1/\delta)}{n}} + C_2(d,R)\cdot\frac{\log(1/\delta)}{n}, \tag{49}$$

where

$$C_1(d,R) = O\left(d^{3/2}R^2\right),\ C_2(d,R) = O\left(R^2\right). \tag{50}$$

Proof of Lemma 6. Note that Gemini 3.1 Pro was used as an aid in proving this lemma, with significant direction and editing by the authors.

Let $\mathcal{N}(\theta_n, \Sigma_n)$ denote the Gaussian best fit to $\hat{\mu}_n$, $\mathcal{N}(\theta, \Sigma)$ denote the Gaussian best fit to μ, and $\mathcal{N}(\theta_{\mathrm{ref}}, \Sigma_{\mathrm{ref}})$ denote the Gaussian best fit to μ_{ref}. We therefore have

$$\begin{aligned}
&W_2^2(\mathcal{N}(\theta_n,\Sigma_n),\mathcal{N}(\theta_{\mathrm{ref}},\Sigma_{\mathrm{ref}})) - W_2^2(\mathcal{N}(\theta,\Sigma),\mathcal{N}(\theta_{\mathrm{ref}},\Sigma_{\mathrm{ref}})) \\
&\quad= \left|\left(\|\theta_n-\theta_{\mathrm{ref}}\|_2^2 + \mathrm{tr}\left(\Sigma_n+\Sigma_{\mathrm{ref}}-2(\Sigma_n\Sigma_{\mathrm{ref}})^{\frac{1}{2}}\right)\right) - \left(\|\theta-\theta_{\mathrm{ref}}\|_2^2 + \mathrm{tr}\left(\Sigma+\Sigma_{\mathrm{ref}}-2(\Sigma\Sigma_{\mathrm{ref}})^{\frac{1}{2}}\right)\right)\right| \\
&\quad\le \left|\|\theta_n-\theta_{\mathrm{ref}}\|_2^2 - \|\theta-\theta_{\mathrm{ref}}\|_2^2\right| + |\mathrm{tr}(\Sigma_n-\Sigma_{\mathrm{ref}})| + 2\left|\mathrm{tr}\left((\Sigma_n\Sigma_{\mathrm{ref}})^{\frac{1}{2}}\right) - \mathrm{tr}\left((\Sigma\Sigma_{\mathrm{ref}})^{\frac{1}{2}}\right)\right| \\
&\quad:= \mathrm{(I)} + \mathrm{(II)} + \mathrm{(III)}.
\end{aligned}$$

We now prove that (I), (II), (III) each concentrate at a rate of $O(1/\sqrt{n})$; summing the resulting bounds and combining with a Union Bound then proves Lemma 6.

Upper bounding (I): By the difference of squares formula and Triangle Inequality,

$$\begin{aligned}
\mathrm{(I)} &\le \left|\|\theta_n-\theta_{\mathrm{ref}}\|_2 - \|\theta-\theta_{\mathrm{ref}}\|_2\right| \cdot \left|\|\theta_n-\theta_{\mathrm{ref}}\|_2 + \|\theta-\theta_{\mathrm{ref}}\|_2\right| \\
&\le \|\theta_n-\theta\| \cdot \left|\|\theta_n-\theta_{\mathrm{ref}}\|_2 + \|\theta-\theta_{\mathrm{ref}}\|_2\right|.
\end{aligned}$$

Now, by standard results on concentration of the mean (we use Equation 1.1 of Hopkins (2020)), we have

$$\Pr\left(\|\theta_n-\theta\|_2 > \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}\log(3/\delta)}{n}}\right) \le \delta/3.$$

Observe that $\mathrm{tr}(\Sigma) = \mathbb{E}\left[\|\mu-\theta\|_2^2\right] \le O(R^2)$. In addition, since Σ is PSD, we have

$$\|\Sigma\|_{\mathrm{op}} \le \mathrm{tr}(\Sigma) \le R^2.$$

Combining with the above display, with probability at least $1-\delta/3$,

$$\|\theta_n-\theta\|_2 \le \sqrt{\frac{R^2}{n}} + \sqrt{\frac{2R^2\log(3/\delta)}{n}} \le \sqrt{\frac{8R^2\log(3/\delta)}{n}}.$$

Now, since μ_{ref} and μ are supported on points with L^2 norm at most R, we have

$$\|\theta_n-\theta_{\mathrm{ref}}\|_2 + \|\theta-\theta_{\mathrm{ref}}\|_2 \le \|\theta_n-\theta\|_2 + 2\|\theta-\theta_{\mathrm{ref}}\|_2 \le \|\theta_n-\theta\|_2 + 4R.$$

Therefore, with probability at least $1-\delta/3$, we have

$$\mathrm{(I)} \le \sqrt{\frac{8R^2\log(3/\delta)}{n}}\left(\sqrt{\frac{8R^2\log(3/\delta)}{n}} + 4R\right) \le O\left(R^2\left\{\sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}\right\}\right). \tag{51}$$

Upper bounding (II): Let $X \sim \mu$ be a random vector, and $X_i \sim \mu$ be n i.i.d. draws from μ. By the linearity of the trace operator and the cyclic property $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, the trace of the true covariance matrix can be expressed as the expected squared ℓ_2-norm of the random vector X:

$$\begin{aligned}
\mathrm{tr}(\Sigma) &= \mathrm{tr}\left(\mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^T\right) \\
&= \mathbb{E}\left[\mathrm{tr}(XX^\top)\right] - \mathbb{E}[X^T]\mathbb{E}[X]
\end{aligned}$$

$$= \mathbb{E}\left[\|X\|_2^2\right] - \|\mathbb{E}[X]\|_2^2.$$

Similarly, the trace of the empirical covariance matrix Σ_n simplifies to the empirical average of the squared ℓ_2-norms of the samples:

$$\begin{aligned}\operatorname{tr}(\Sigma_n) &= \operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^\top \left(\frac{1}{n}\sum_{i=1}^n X_i\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^n \operatorname{tr}(X_i X_i^\top) - \left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_2^2 \\ &= \frac{1}{n}\sum_{i=1}^n \|X_i\|_2^2 - \left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_2^2.\end{aligned}$$

Let us define a sequence of scalar random variables $Z_i = \|X_i\|_2^2$ for $i \in \{1, \dots, n\}$. Notice $\mathbb{E}[X] = \theta$ and that $\frac{1}{n}\sum_{i=1}^n X_i, \theta_n$ have identical law. Consequently by our work from the upper bound in (I), and as the X_i are i.i.d., we have with probability at least $1 - \delta/3$,

$$\begin{aligned}\text{(II)} &\le \frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) + \left(\|\theta_n\|_2^2 - \|\theta\|^2\right) \\ &\le \frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) + \left(\|\theta_n\| + \|\theta\|\right) \cdot \|\theta_n - \theta\| \\ &\le \frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) + 2R \cdot O\left(\sqrt{\frac{8R^2 \log(3/\delta)}{n}}\right), \end{aligned} \tag{52}$$

where we use that μ, μ_{ref} are supported on $\mathbb{B}(0, R)$.

As each $\|X_i\|_2 \le R$, it follows that $Z_i \in [0, R^2]$. Moreover the Z_i are independent. Thus by Hoeffding's inequality, for any $t > 0$, we have:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])\right| \ge t\right) \le 2\exp\left(-\frac{2nt^2}{(R^2 - 0)^2}\right) = 2\exp\left(-\frac{2nt^2}{R^4}\right). \tag{53}$$

We equate the RHS of the above to $\delta/3$ and therefore set

$$t = R^2\sqrt{\frac{\log(6/\delta)}{2n}}. \tag{54}$$

Therefore combining with (52), with probability at least $1 - 2\delta/3$, we conclude that

$$\text{(II)} = |\operatorname{tr}(\Sigma_n) - \operatorname{tr}(\Sigma)| \le O\left(R^2\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{55}$$

Upper bounding (III): We turn to upper bounding (III). Obtaining a $1/\sqrt{n}$ rate for this quantity is non-trivial; direct arguments give a $1/n^{1/4}$ convergence rate. In the 1d case, this amounts to showing that the empirical standard deviation converges with high probability at a $\frac{1}{\sqrt{n}}$ rate to the true standard deviation, and we begin the proof in this case.

Lemma 7. *Let $X_1, \dots, X_n \sim \mu$ be i.i.d. random variables drawn from a distribution μ supported on $[-R, R]$ with true standard deviation σ. Let $\widehat{\sigma}$ be the empirical sample standard deviation. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that:*

$$|\widehat{\sigma} - \sigma| \le 9R\sqrt{\frac{\log(4/\delta)}{n}}. \tag{56}$$

Proof. Let $\widehat{\theta} := \frac{1}{n}\sum X_i$ be the empirical mean and $\theta := \mathbb{E}[X]$ be the true mean. We wish to bound the fully empirical variance, $\widehat{\sigma}^2 = \frac{1}{n}\sum (X_i - \widehat{\theta})^2$. By algebraic expansion, we have

$$\widehat{\sigma}^2 = \frac{1}{n}\sum (X_i - \theta)^2 - (\widehat{\theta} - \theta)^2 \implies \left|\widehat{\sigma}^2 - \sigma^2\right| \le \left|\frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 - \sigma^2\right| + (\widehat{\theta} - \theta)^2. \tag{57}$$

We bound the two terms on the right-hand side individually.

1. First term: Observe that $\mathbb{E}\big[(X_i-\theta)^2-\sigma^2\big]=\mathbb{E}\big[\mathbb{V}(X_i)-\sigma^2\big]=\mathbb{E}\big[\mathbb{E}[X_i^2]-\mathbb{E}[X_i]^2\big]-\sigma^2=\sigma^2-\sigma^2=0$. Since $\mathcal{D}$ is supported on $[-R,R]$, $(X_i-\theta)^2\le 4R^2$. Applying Bennett's inequality (Maurer & Pontil, 2009), we obtain with probability at least $1-\delta/2$:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\theta)^2-\sigma^2\right|\le 4\sigma R\sqrt{\frac{2\log(2/\delta)}{n}}+\frac{8R^2}{3}\cdot\frac{\log(2/\delta)}{n}.$$

where we bounded variance in the following way: The variance of Y_i is identical to the variance of the uncentered squared deviation $(X_i-\theta)^2$. We can bound this by the second uncentered moment of $(X_i-\theta)^2$, which is the fourth central moment of X_i:

$$\begin{aligned}\mathrm{Var}(Y_i)&=\mathrm{Var}\left((X_i-\theta)^2\right)\\&=\mathbb{E}\left[(X_i-\theta)^4\right]-\left(\mathbb{E}\left[(X_i-\theta)^2\right]\right)^2\\&\le\mathbb{E}\left[(X_i-\theta)^4\right].\end{aligned}$$

Because the support of the distribution D is contained entirely within $[-R,R]$, the maximum deviation from the true mean θ is almost surely bounded by $2R$. Consequently, the squared deviation satisfies $(X_i-\theta)^2\le 4R^2$ with probability 1. We can factor this worst-case deterministic bound out of the expectation to relate the fourth moment back to the true variance σ^2:

$$\begin{aligned}\mathbb{E}\left[(X_i-\theta)^4\right]&=\mathbb{E}\left[(X_i-\theta)^2(X_i-\theta)^2\right]\\&\le\mathbb{E}\left[4R^2(X_i-\theta)^2\right]\\&=4R^2\mathbb{E}\left[(X_i-\theta)^2\right]\\&=4R^2\sigma^2.\end{aligned}$$

Thus, we establish the uniform variance bound $\mathrm{Var}(Y_i)\le 4R^2\sigma^2$, which we substitute directly into the standard formulation of Bernstein's inequality.

2. Second term: Applying Hoeffding's inequality (Hoeffding, 1963) to the empirical mean, since $\mathcal{D}$ is suppoted on $[-R,R]$ we obtain with probability at least $1-\delta/2$,

$$(\widehat{\theta}-\theta)^2\le 2R^2\cdot\frac{\log(4/\delta)}{n}.$$

Taking a union bound over these two events and combining with (57), we obtain with probability at least $1-\delta$,

$$\left|\widehat{\sigma}^2-\sigma^2\right|\le 4\sigma R\sqrt{\frac{\log(2/\delta)}{n}}+5R^2\cdot\frac{\log(4/\delta)}{n},\tag{58}$$

However, note the right hand side above depends on σ. To establish the desired result (56), we condition on the event from (58), which has probability at least $1-\delta$. Under this event where the inequality (58) holds, we break into two cases based on the value of σ compared to the threshold $\gamma:=R\sqrt{\frac{\log(4/\delta)}{n}}$.

- **Large Variance ($\sigma\ge\gamma$):** Since $\widehat{\sigma}\ge 0$ and $\sigma>0$, we have the bound $|\widehat{\sigma}-\sigma|\le\frac{|\widehat{\sigma}^2-\sigma^2|}{\widehat{\sigma}+\sigma}\le\frac{|\widehat{\sigma}^2-\sigma^2|}{\sigma}$. Dividing both sides of (58) by σ, noting $\sigma\ge\gamma$, and recalling the definition of γ gives

$$|\widehat{\sigma}-\sigma|\le 4R\sqrt{\frac{\log(2/\delta)}{n}}+5\cdot\frac{R^2}{\sigma}\frac{\log(4/\delta)}{n}\le 4\gamma+5\cdot\frac{R^2\cdot\frac{\log(4/\delta)}{n}}{R\sqrt{\frac{\log(4/\delta)}{n}}}=(4+5)\gamma=9\gamma.$$

- **Small Variance ($\sigma<\gamma$):** Let $f(\sigma):=4\sigma\gamma+5\gamma^2$. Since γ does not depend on σ, $f(\sigma)$ is non-decreasing with respect to σ. As $\sigma<\gamma$, we have by (58),

$$|\widehat{\sigma}^2-\sigma^2|\le f(\sigma)\le f(\gamma)=(4+5)\gamma^2=9\gamma^2.$$

Note for any $a,b\ge 0$, $|a-b|\le\sqrt{|a^2-b^2|}$. Combining this fact with the above display establishes (56) in this case.

We have thus established (56) in both cases, proving Lemma 7. □

We now leverage Lemma 7 to establish a similar result in the multivariate case.

Lemma 8. *Let μ be a distribution over $\mathbb{R}^d$ supported almost surely on $\{x : \|x\|_2 \le R\}$. Consider n i.i.d. samples $X_1, \dots, X_n$ from μ, and let $\Sigma := \mathbb{E}[XX^\top]$ for $X \sim \mu$ and $\Sigma_n := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. Then, for any fixed PSD matrix Σ_{ref} with operator norm $\|\Sigma_{\mathrm{ref}}\|_{\mathrm{op}} \le M$, we have for all $\delta \in (0,1)$,*

$$\left|\mathrm{tr}\Big((\Sigma_{\mathrm{ref}}^{1/2}\Sigma\Sigma_{\mathrm{ref}}^{1/2})^{1/2}\Big) - \mathrm{tr}\Big((\Sigma_{\mathrm{ref}}^{1/2}\Sigma_n\Sigma_{\mathrm{ref}}^{1/2})^{1/2}\Big)\right| \le 18dR\sqrt{M}\sqrt{\frac{d\log(9)+\log(4/\delta)}{n}}. \tag{59}$$

Proof. Consider any orthonormal basis $\{w_k\}$ of $\mathbb{R}^d$. By the Cauchy-Schwarz Inequality, for any vectors $u, v \in \mathbb{R}^d$, we have $u^\top v \le \|u\|_2\|v\|_2$. Letting $u = w_k$ and $v = X^{1/2}w_k$, we obtain:

$$w_k^\top(X^{1/2}w_k) \le \|w_k\|_2\|X^{1/2}w_k\|_2. \tag{60}$$

Since $\{w_k\}$ is an orthonormal basis, each $\|w_k\|_2 = 1$. Furthermore, because X is PSD, its principal square root $X^{1/2}$ is symmetric, that is $(X^{1/2})^\top = X^{1/2}$. We can compute the squared Euclidean norm of $X^{1/2}w_k$ as:

$$\|X^{1/2}w_k\|_2^2 = (X^{1/2}w_k)^\top(X^{1/2}w_k) = w_k^\top(X^{1/2})^\top X^{1/2}w_k = w_k^\top X w_k. \tag{61}$$

Taking the square root yields $\|X^{1/2}w_k\|_2 = \sqrt{w_k^\top X w_k}$. Combining the above displays thus gives

$$w_k^\top X^{1/2}w_k \le \|X^{1/2}w_k\|_2 \le \sqrt{w_k^\top X w_k}. \tag{62}$$

Summing this inequality over all $k \in [d]$, we obtain

$$\mathrm{tr}(X^{1/2}) = \sum_{k=1}^d w_k^\top X^{1/2}w_k \le \sum_{k=1}^d \sqrt{w_k^\top X w_k}. \tag{63}$$

In addition, one can see that equality holds in all the steps above when $\{w_k\}_{k=1}^d$ is the eigenbasis of X, so if $\{w_k\}_{k=1}^d$ is the eigenbasis of X, then

$$\mathrm{tr}(X^{1/2}) = \sum_{k=1}^d \sqrt{w_k^\top X w_k}. \tag{64}$$

We now let

$$X_\Sigma = (\Sigma_{\mathrm{ref}}^{1/2}\Sigma\Sigma_{\mathrm{ref}}^{1/2}) \quad , \quad X_{\Sigma_n} = (\Sigma_{\mathrm{ref}}^{1/2}\Sigma_n\Sigma_{\mathrm{ref}}^{1/2}). \tag{65}$$

Let $U = \{u_k\}_{k=1}^d$ denote the eigenbasis of X_Σ and $V = \{v_k\}_{k=1}^d$ denote the eigenbasis of X_{Σ_n}. Applying (63) with V, we have

$$\mathrm{tr}(X_\Sigma^{1/2}) \le \sum_{k=1}^d \sqrt{v_k^\top X_\Sigma v_k}. \tag{66}$$

Applying (64) with V, we have

$$\mathrm{tr}(X_{\Sigma_n}^{1/2}) = \sum_{k=1}^d \sqrt{v_k^\top X_{\Sigma_n} v_k}. \tag{67}$$

Subtracting (66) from (67),

$$\mathrm{tr}(X_\Sigma^{1/2}) - \mathrm{tr}(X_{\Sigma_n}^{1/2}) \le \sum_{k=1}^d \left(\sqrt{v_k^\top X_{\Sigma_n} v_k} - \sqrt{v_k^\top X_\Sigma v_k}\right).$$

Similarly, using the basis U,

$$\mathrm{tr}(X_{\Sigma_n}^{1/2}) - \mathrm{tr}(X_\Sigma^{1/2}) \le \sum_{k=1}^d \left(\sqrt{u_k^\top X_{\Sigma_n} u_k} - \sqrt{u_k^\top X_\Sigma u_k}\right).$$

Thus, we have

$$\left|\mathrm{tr}(X_{\Sigma_n}^{1/2}) - \mathrm{tr}(X_{\Sigma}^{1/2})\right| \le \max_{W\in\{U,V\}} \sum_{k=1}^{d} \left|\sqrt{w_k^\top X_{\Sigma_n} w_k} - \sqrt{w_k^\top X_{\Sigma} w_k}\right|. \tag{68}$$

Next, we analyze the individual terms within (68) for a specific unit vector $w \in \mathcal{S}^{d-1}$. We expand the quadratic form involving X_{Σ_n}:

$$w^\top X_{\Sigma_n} w = w^\top \Sigma_{\mathrm{ref}}^{1/2} \Sigma_n \Sigma_{\mathrm{ref}}^{1/2} w = (\Sigma_{\mathrm{ref}}^{1/2} w)^\top \Sigma_n (\Sigma_{\mathrm{ref}}^{1/2} w). \tag{69}$$

Define the transformed vector $\tilde{w} = \Sigma_{\mathrm{ref}}^{1/2} w$. Because w is a unit vector and the operator norm of Σ_{ref} is bounded by M, the Euclidean norm of this transformed vector is bounded as follows:

$$\|\tilde{w}\|_2 \le \|\Sigma_{\mathrm{ref}}^{1/2}\|_{\mathrm{op}} \|w\|_2 \le \sqrt{M}. \tag{70}$$

Substituting the explicit formula for the empirical covariance matrix Σ_n, we have

$$\tilde{w}^\top \Sigma_n \tilde{w} = \tilde{w}^\top \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\right) \tilde{w} = \frac{1}{n}\sum_{i=1}^{n} (\tilde{w}^\top X_i)^2. \tag{71}$$

This reveals that $w^\top X_{\Sigma_n} w$ is exactly the empirical variance $\widehat{\sigma}_{\tilde{w}}^2$ of the 1d random variable $Y_i := \tilde{w}^\top X_i$. Similarly, $w^\top X_{\Sigma} w$ is the true expected variance $\sigma_{\tilde{w}}^2 = \mathbb{V}(\tilde{w}^\top X)$. Therefore, we have

$$\left|\sqrt{w^\top X_{\Sigma_n} w} - \sqrt{w^\top X_{\Sigma} w}\right| = |\widehat{\sigma}_{\tilde{w}} - \sigma_{\tilde{w}}|. \tag{72}$$

To apply Lemma 7 on 1d standard deviation concentration, we must establish an absolute upper bound on the random variable y_i. By the Cauchy-Schwarz Inequality and our prior norm bounds:

$$|Y_i| = |\tilde{w}^\top X_i| \le \|\tilde{w}\|_2 \|X_i\|_2 \le \sqrt{M} R, \tag{73}$$

as μ is supported on $\mathbb{B}(0, R)$. Thus, Y_i is strictly supported on the interval $[-R\sqrt{M}, R\sqrt{M}]$. By Lemma 7, for any *fixed* vector w and for all $\delta' \in (0, 1)$, with probability at least $1 - \delta'$:

$$|\widehat{\sigma}_{\tilde{w}} - \sigma_{\tilde{w}}| \le 9R\sqrt{M}\sqrt{\frac{\log(4/\delta')}{n}}. \tag{74}$$

Because the eigenbasis V is depends on the empirical data Σ_n, we must guarantee that this concentration bound holds uniformly over all possible directions $w \in \mathcal{S}^{d-1}$. We construct an ε-net $\mathcal{N}_\varepsilon$ over the unit sphere with $\varepsilon = 1/4$. By a volumetric argument, the cardinality of this net is bounded by $|\mathcal{N}_{1/4}| \le (1 + \frac{2}{1/4})^d = 9^d$; see e.g. Vershynin (2018). Applying a Union Bound over all points within $\mathcal{N}_{1/4}$ and setting $\delta' = \delta/9^d$ in (74), we have for all w in the net, we have with probability at least

$$\left|\sqrt{w^\top X_{\Sigma_n} w} - \sqrt{w^\top X_{\Sigma} w}\right| = |\widehat{\sigma}_{\tilde{w}} - \sigma_{\tilde{w}}| \le 9R\sqrt{M}\sqrt{\frac{d\log(9) + \log(4/\delta)}{n}}. \tag{75}$$

. Here we used that $\log(4/\delta') = \log(4 \cdot 9^d/\delta) = d\log(9) + \log(4/\delta)$.

To extend this bound from the ε-net to the continuous sphere, let $M_X = X_{\Sigma_n} - X_{\Sigma}$. Consequently we have

$$\|M_X\|_{\mathrm{op}} = \sup_{w\in\mathcal{S}^{d-1}} |w^\top M_X w|.$$

Let $w^\star$ be the unit vector achieving this supremum. By the definition of the net, there exists $v \in \mathcal{N}_{1/4}$ such that $\|w^\star - v\|_2 \le 1/4$. Expanding the bilinear difference and applying the triangle inequality yields:

$$\begin{aligned}
|(w^\star)^\top M_X w^\star - v^\top M_X v| &\le |(w^\star)^\top M_X (w^\star - v)| + |(w^\star - v)^\top M_X v| \\
&\le \|M_X\|_{\mathrm{op}} \|w^\star\|_2 \|w^\star - v\|_2 + \|M_X\|_{\mathrm{op}} \|w^\star - v\|_2 \|v\|_2 \\
&\le 2\left(\frac{1}{4}\right)\|M_X\|_{\mathrm{op}} = \frac{1}{2}\|M_X\|_{\mathrm{op}}.
\end{aligned}$$

By the Triangle Inequality,

$$\|M_X\|_{\mathrm{op}} = |(w^\star)^\top M_X w^\star| \le |v^\top M_X v| + \frac{1}{2}\|M_X\|_{\mathrm{op}} \implies \|M_X\|_{\mathrm{op}} \le 2 \max_{v \in \mathcal{N}_{1/4}} |v^\top M_X v|.$$

Combining with (75) gives that with probability at least $1-\delta$, for all $w \in \mathcal{S}^{d-1}$,

$$\left|\sqrt{w^\top X_{\Sigma_n} w} - \sqrt{w^\top X_\Sigma w}\right| \le 18R\sqrt{M}\sqrt{\frac{d\log(9)+\log(4/\delta)}{n}}. \tag{76}$$

We now combine with (68) and sum (76) across the d basis vectors in either U or V (whichever is the argmax in (68)) to obtain

$$\left|\mathrm{tr}(X_{\Sigma_n}^{1/2}) - \mathrm{tr}(X_\Sigma^{1/2})\right| \le 18dR\sqrt{M}\sqrt{\frac{d\log(9)+\log(4/\delta)}{n}}. \tag{77}$$

This completes the proof of Lemma 8. □

We now complete the proof of Lemma 6. By Lemma 8 and commutativity of trace, we obtain with probability at least $1-\delta/3$,

$$\begin{aligned}\left|\mathrm{tr}\Big((\Sigma_{\mathrm{ref}}^{1/2}\Sigma\Sigma_{\mathrm{ref}}^{1/2})^{1/2}\Big) - \mathrm{tr}\Big((\Sigma_{\mathrm{ref}}^{1/2}\Sigma_n\Sigma_{\mathrm{ref}}^{1/2})^{1/2}\Big)\right| &= \left|\mathrm{tr}\left((\Sigma_n\Sigma_{\mathrm{ref}})^{\frac{1}{2}}\right) - \mathrm{tr}\left((\Sigma\Sigma_{\mathrm{ref}})^{\frac{1}{2}}\right)\right| \\ &\le 18dR^2\sqrt{\frac{d\log(9)+\log(12/\delta)}{n}}. \end{aligned} \tag{78}$$

where we have used $M = R^2$ since the operator norm of Σ_{ref} is upper bounded by R^2, as discussed earlier. Lemma 6 now follows by a Union Bound over the high-probability upper bounds (51), (55), (78) on (I), (II), (III) respectively. □

Lemma 9 (Convexity of FID loss over distributions)**.** *Consider any distribution μ in the mixture form $\mu = \sum_{i=1}^n \lambda_i \mu_i$ for $\sum_{i=1}^n \lambda_i = 1$ and $\lambda_i \ge 0$. Then for any μ_{ref},*

$$\mathrm{FID}(\mu, \mu_{\mathrm{ref}}) \le \sum_{i=1}^n \lambda_i \mathrm{FID}(\mu_i, \mu_{\mathrm{ref}}).$$

Proof. Letting $\hat{\mu}_{\mathrm{ref}}$ denote the Gaussian with mean equaling the empirical mean of μ_{ref} and covariance equaling the empirical covariance of μ_{ref}, we observe that $\mathrm{FID}(\mu, \mu_{\mathrm{ref}}) = \mathrm{FID}(\mu, \hat{\mu}_{\mathrm{ref}})$, $\mathrm{FID}(\mu_i, \mu_{\mathrm{ref}}) = \mathrm{FID}(\mu_i, \hat{\mu}_{\mathrm{ref}})$. Consequently we may suppose without loss of generality that μ_{ref} is Gaussian.

Let us define

$$F_i := \mathrm{FID}(\mu_i, \mu_{\mathrm{ref}}) \text{ for all } i = 1, \dots, K \quad \text{and} \quad F := \mathrm{FID}(\mu, \mu_{\mathrm{ref}}).$$

Now, define μ_i' as the Gaussian best fit to μ_i, for $i = 1, \dots, K$. In particular, $\mu_i' = \mathcal{N}(\theta_i, \Sigma_i)$ where θ_i and Σ_i are the mean and covariance of μ_i. Define $\mu' = \sum_{i=1}^K \lambda_i \mu_i'$. In addition, let us define θ and Σ as the mean and covariance of μ. We have that

$$\theta = \sum_{i=1}^K \lambda_i \theta_i \quad , \quad \Sigma = \sum_{i=1}^K \lambda_i \Big(\Sigma_i + (\theta_i - \theta)(\theta_i - \theta)^\top\Big),$$

and also that θ is the mean and Σ is the covariance of μ'.

Let us define

$$\begin{aligned} F_i' &:= \mathrm{FID}(\mu_i', \mu_{\mathrm{ref}}) \text{ for all } i = 1, \dots, K \quad \text{and} \quad F' := \mathrm{FID}(\mu', \mu_{\mathrm{ref}}), \\ W_i' &:= W_2^2(\mu_i', \mu_{\mathrm{ref}}) \text{ for all } i = 1, \dots, K \quad \text{and} \quad W' := W_2^2(\mu', \mu_{\mathrm{ref}}), \end{aligned}$$

where W_2 denotes the 2-Wasserstein distance.

Since all of the μ_i', μ are Gaussian and μ_{ref} is Gaussian (as we showed we can assume without loss of generality), $W_i' = F_i'$ for $i = 1, \dots, K$. By convexity of W_2^2 (Panaretos & Zemel, 2020), we thus have

$$W' \le \sum_{i=1}^K \lambda_i W_i' = \sum_{i=1}^K \lambda_i F_i'.$$

In addition, by Gelbrich's bound (Gelbrich, 1990), we have $F' \leq W'$, thus combining with the above gives

$$F' \leq W' \leq \sum_{i=1}^{K} \lambda_i F_i'.$$

Since the FID loss depends only on the mean and covariance of the reference and sample distribution, we have $F_i' = F_i$ for $i = 1, \ldots, K$. Furthermore, since μ and μ' have the same mean and covariance, we also have $F' = F$. Therefore we obtain, combining with the above,

$$F = F' \leq \sum_{i=1}^{K} \lambda_i F_i' = \sum_{i=1}^{K} \lambda_i F_i,$$

which is the desired bound. □

Lemma 10. *Consider two distributions μ_1, μ_2 supported on $\mathbb{B}(0, R)$ in $\mathbb{R}^d$, with empirical mean and covariance θ_1, Σ_1 and θ_2, Σ_2 respectively. Letting $\mathcal{N}_1$ denote the Gaussian $N(\theta_1, \Sigma_1)$ and similarly defining $\mathcal{N}_2$, we have*

$$W_2(\mathcal{N}_1, \mathcal{N}_2)^2 \leq O(R^2 d).$$

Proof. We note that

$$W_2(\mathcal{N}_1, \mathcal{N}_2)^2 = \|\theta_1 - \theta_2\|^2 + \mathrm{tr}\big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}\big).$$

Observe that $\theta_1, \theta_2 \in B(0, R)$. Consequently $\|\theta_1 - \theta_2\|^2 \leq 4R^2$. Next we have $\Sigma_1, \Sigma_2 \preceq R^2 I_d$. Thus as $\Sigma_1\Sigma_2$ is PSD, we have

$$\mathrm{tr}\big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}\big) \leq \mathrm{tr}(\Sigma_1 + \Sigma_2) \leq 2R^2 d.$$

Combining these bounds implies the Lemma. □