The Capabilities and Limitations of Weak-to-Strong Generalization: Generalization and Calibration

Anonymous ACL submission

Abstract

Weak-to-strong generalization, where weakly supervised strong models outperform their weaker teachers, offers a promising approach to aligning superhuman models with human values. To deepen the understanding of this approach, we provide theoretical insights into its capabilities and limitations. First, in the classification setting, we establish upper and lower generalization error bounds for the strong model, identifying the primary limitations as stemming from the weak model's generalization error and the optimization objective itself. Additionally, we derive lower and upper bounds on the calibration error of the strong model. These theoretical bounds reveal two critical insights: (1) the weak model should demonstrate strong generalization performance and maintain well-calibrated predictions, and (2) the strong model's training process must strike a careful balance, as excessive optimization may lead to overfitting to the weak supervision. Finally, in the regression setting, we theoretically extend the work of Charikar et al. (2024) to a loss function based on KL divergence, offering guarantees that the strong student can outperform its weak teacher by at least the magnitude of their disagreement. The theory is validated through synthetic experiments.

1 Introduction

Human supervision (Ouyang et al., 2022; Bai et al., 2022a) plays a crucial role in building both effective and safe artificial intelligence systems (Achiam et al., 2023; Touvron et al., 2023). However, as future superhuman models exhibit increasingly complex behaviors, reliable human oversight becomes increasingly challenging (OpenAI, 2024).

To tackle this issue, the Weak-To-Strong Generalization (W2SG) paradigm (Burns et al., 2023) is proposed. It finds that, strong pre-trained language models, when fine-tuned using labels produced by weaker models, consistently achieve better performance than their weak supervisors. This intriguing phenomenon has not only driven the development of diverse alignment algorithms (Zhu et al., 2025; Liu and Alahi, 2024), but also inspired efforts (Pawelczyk et al., 2024; Yang et al., 2024; Guo et al., 2024) to extend the concept to other tasks. However, despite its empirical success, the theoretical foundations of W2SG remain underdeveloped. Although several elegant theoretical frameworks (Lang et al., 2024; Somerstep et al., 2024; Wu and Sahai, 2025; Charikar et al., 2024) are proposed, a universal framework is still lacking to address fundamental questions, such as: *What is the optimal generalization performance a strong model can achieve after W2SG? What other factors are influenced by W2SG?*

To answer these questions, we provide a theoretical analysis of W2SG, shedding lights on its capabilities and limitations. Firstly, in classification tasks, our theoretical analysis of lower and upper generalization bounds under KL divergence loss reveals that the strong model's performance is fundamentally determined by two key factors: (1) the disagreement between strong and weak models, which serves as the minimization objective in W2SG, and (2) the weak model's performance. These findings suggest that (1) achieving the minimal optimization objective in W2SG limits the strong model's ability to significantly outperform its weak supervisor, and (2) selecting a stronger weak model can enhance the performance of the strong model. Secondly, we investigate how strong model's calibration—the property that a model's predicted confidence aligns with its actual accuracy (Guo et al., 2017; Kumar et al., 2019)—is affected in the W2SG framework. Our theoretical bounds reveal that the calibration of the strong model depends on both the calibration of the weak model and the disagreement between the two models. The theory highlights the importance of avoiding a poorly-calibrated weak model and an overfitted strong model. The above theoretical analysis is validated using GPT-2 series (Radford et al.,

2019) and Pythia series (Biderman et al., 2023).

In addition to classification setting, we also consider the regression problem. In particular, we build on Charikar et al. (2024) by extending their analysis of squared loss to output distribution divergence, a measure of the difference between two models' output distributions. In this setting, the model outputs are normalized to form valid probability distributions over all input data, and the output distribution divergence between two models is defined as the KL divergence of their respective output distributions. We recover the findings from Charikar et al. (2024) and show that the strong model's generalization error is provably smaller than the weak model's, with the gap no less than the W2SG minimization objective—namely, the strong model's error on the weak labels. We conduct synthetic experiments to support our theoretical insights.

2 Related Work

In this section, we introduce AI alignment and W2SG. Additional related work including teacher-student learning paradigm, weakly-supervised learning, calibration and information-theoretic analysis is provided in Appendix A.

AI alignment. AI alignment (Ji et al., 2023; Shen et al., 2023) aims to ensure AI systems act in accordance with human values. A popular approach to achieve this goal is fine-tuning models on human-annotated data, such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a) and Direct Preference Optimization (DPO) (Rafailov et al., 2024). However, this alignment paradigm faces significant challenges: human oversight becomes insufficient as AI surpasses human capabilities (Kim et al., 2024), and obtaining scalable, high-quality human feedback remains difficult (Casper et al., 2023). These challenges highlight the critical need to align superhuman AI systems (OpenAI, 2024). In contrast to these approaches, our work explores W2SG, which leverages weak models as the teacher to achieve the alignment goal.

Weak-to-strong generalization. To explore the effect of weak models to supervise strong models, Burns et al. (2023) first find that strong models supervised by weak models can exhibit better performance on corresponding tasks than their weak supervisors, indicating the possibility of stimulating greater power from super models under weak supervisions. There are also algorithms (Zhu et al., 2025; Agrawal et al., 2024; Sang et al., 2024; Guo and Yang, 2024) and empirical analysis (Yang et al., 2025; Ye et al., 2024) for it. However, only a limited number of theoretical studies have been conducted on this topic. Lang et al. (2024) analyzes it by introducing theoretical bounds that account for pseudolabel correction and coverage expansion. Somerstep et al. (2024) frame W2SG as a transfer learning problem, revealing limitations of fine-tuning on weak labels. Wu and Sahai (2025) study linear models under a spiked covariance setting and derive asymptotic bounds. Charikar et al. (2024) take a convex-theoretic approach in regression, quantifying performance improvements under squared loss via the misfit error between weak and strong models. The work most closely related to ours is Charikar et al. (2024), which primarily focuses on squared loss in regression. In contrast, we consider KL divergence-like losses, including KL divergence for classification and output distribution divergence for regression. Furthermore, while they focuses on establishing upper bounds, our study incorporates both upper and lower bounds as well as calibration analysis through experiments on language models, providing a more comprehensive understanding of the fundamental capabilities and limitations of W2SG.

3 Preliminaries

3.1 Classification and Regression

We examine two problem settings. In the first case, we consider classification tasks using KL divergence as the loss function. Minimizing this loss is equivalent to minimizing cross-entropy loss, which is widely used in the W2SG literature (Burns et al., 2023). In the second case, we focus on regression tasks, employing the KL divergence between the predictions of two models as the loss function. The model outputs over the entire data domain are normalized to form probability distributions. This approach is an extension of previous result (Charikar et al., 2024) on squared loss, and provides an intuitive framework for understanding W2SG.

Given the data distribution $\mathcal{P}$, data domain $\mathcal{X}$ and output domain $\mathcal{Y}$, let $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$. Consider the difference $d_{\mathcal{P}}$ and empirical difference $\hat{d}_{\mathcal{P}}$ between two models, where $d_{\mathcal{P}}, \hat{d}_{\mathcal{P}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_0^+$. We define the below two settings:

Setting 1: KL divergence loss. Firstly, we consider a k-classification problem. Given the data domain $\mathcal{X} \subseteq \mathbb{R}^d$ and output domain $\mathcal{Y} \subseteq \mathbb{R}^k$. Con-

sider the model with the softmax module, i.e., $\forall y = (y_1, \cdots, y_k)^T \in \mathcal{Y}$, there holds $\sum_{i=1}^k y_i = 1$ and $0 < y_i \leq 1$. Given two models $f, g \in \mathcal{F}$, define $d_{\mathcal{P}}$ and $\hat{d}_{\mathcal{P}}$:

$$d_{\mathcal{P}}(f, g) \triangleq \mathbb{E}_{x \sim \mathcal{P}} \left[\mathrm{D}_{\mathrm{KL}}(f(x) \| g(x)) \right], \quad (1)$$

$$\hat{d}_{\mathcal{P}}(f, g) \triangleq \frac{1}{n} \sum_{j=1}^{n} \mathrm{D}_{\mathrm{KL}}(f(x_j) \| g(x_j)), \quad (2)$$

where the KL divergence $\mathrm{D}_{\mathrm{KL}}(f(x) \| g(x)) = \sum_{i=1}^k [f(x)]_i \log \frac{[f(x)]_i}{[g(x)]_i}$ is the KL divergence between predictions, and $[f(x)]_i, [g(x)]_i$ are elements of $f(x), g(x)$.

Setting 2: Output distribution divergence. While KL divergence loss serve as de facto standard for classification, recent research demonstrates that KL divergence has been effectively employed in regression (Imani and White, 2018; Yang et al., 2021; Günder et al., 2022; Kitazawa, 2025). Therefore, we also consider a regression problem using KL divergence in W2SG. Let the data domain and output domain be $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{y \in \mathbb{R} | 0 < y \leq 1\}$, respectively. In this setting, the outputs of the model for all input data are probability-normalized to ensure they form valid probability distributions. The difference between two models $f, g \in \mathcal{F}$ is then measured as the KL divergence between their corresponding output distributions:

$$d_{\mathcal{P}}(f, g) \triangleq \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx, \quad (3)$$

$$\hat{d}_{\mathcal{P}}(f, g) \triangleq \sum_{i=1}^{n} f(x_i) \log \frac{f(x_i)}{g(x_i)}. \quad (4)$$

3.2 Weak-to-strong Generalization

In the context of W2SG, we focus on the fine-tuning phase after pre-training. Let $h^\star : \mathbb{R}^d \to \mathbb{R}^{d^\star}$ denotes the ground truth representation function, which maps data $x \in \mathcal{X}$ to an ideal, fully enriched representation $h^\star(x)$. The target fine-tuning task, composed with the ground truth representation, is denoted as $f^\star \circ h^\star$, where $f^\star : \mathbb{R}^{d^\star} \to \mathcal{Y}$. The *weak model* learns a mapping $f_w \circ h_w$, where the pre-trained representation $h_w : \mathcal{X} \to \mathbb{R}^{d_w}$ extracts features from the input data, and $f_w : \mathbb{R}^{d_w} \to \mathcal{Y}$ is fine-tuned using supervised data with ground truth labels. The strong model, on the other hand, aims to learn a mapping $f_{sw} \circ h_s$, where $h_s : \mathcal{X} \to \mathbb{R}^{d_s}$ is the representation, and $f_{sw} \in \mathcal{F}_s$ is a task-specific function from a hypothesis class $\mathcal{F}_s : \mathbb{R}^{d_s} \to \mathcal{Y}$. The strong model leverages the representation h_s to improve performance on the fine-tuning task. In the convention setting of AI alignment (Ouyang et al., 2022), the model is learned through human-annotated ground truth data:

$$f_s = \mathrm{argmin}_{f \in \mathcal{F}_s} \, d_{\mathcal{P}}(f^\star \circ h^\star, f \circ h_s). \quad (5)$$

Nevertheless, the acquisition of human-generated data is both costly and time-consuming. To address this challenge, the W2SG framework leverages weak supervision from the weak model's predictions, enabling the strong model to be trained through population risk minimization:

$$f_{sw} = \mathrm{argmin}_{f \in \mathcal{F}_s} \, d_{\mathcal{P}}(f_w \circ h_w, f \circ h_s). \quad (6)$$

In practice, we label n i.i.d. samples using the weak model and minimize the empirical risk:

$$\hat{f}_{sw} = \mathrm{argmin}_{f \in \mathcal{F}_s} \hat{d}_{\mathcal{P}}(f_w \circ h_w, f \circ h_s). \quad (7)$$

Denote the labeling function $F^\star = f^\star \circ h^\star$, strong ceiling model $F_s = f_s \circ h_s$, weak model $F_w = f_w \circ h_w$, and strong models $F_{sw} = f_{sw} \circ h_s$, $\hat{F}_{sw} = \hat{f}_{sw} \circ h_s$, respectively.

4 Universal Results in W2SG

In this section, we consider the classification problem, where $d_{\mathcal{P}}$ is the KL divergence loss defined in Equation (2). We first establish lower and upper generalization error bounds of the strong model in W2SG in Section 4.1. Then the lower and upper calibration error bounds are shown in Section 4.2.

4.1 Generalization Error Bounds

Theorem 4.1 (Proved in Appendix B.1)**.** *Given the data domain $\mathcal{X}$, output domain $\mathcal{Y}$ and models $F_{sw}, F_w, F^\star$ defined above. Then there holds*

$$|d_{\mathcal{P}}(F^\star, F_{sw}) - d_{\mathcal{P}}(F^\star, F_w)| \leq \mathcal{O}\left(\sqrt{d_{\mathcal{P}}(F_{sw}, F_w)}\right),$$

Remark. *The proof can be also extended to the regression setting, which is provided in Appendix B.2.*

Theorem 4.1 provides a quantitative framework for assessing the performance gap between weak model and strong model in W2SG. Specifically, the value of $d_{\mathcal{P}}(F^\star, F_{sw})$ is constrained by two terms: (1) $d_{\mathcal{P}}(F^\star, F_{sw})$, which reflects the performance of the weak model, and (2) $d_{\mathcal{P}}(F_w, F_{sw})$, which is decided by the optimization result in Equation (6) and measures how the strong model learns to imitate the weak supervisor. This result is examined from two complementary perspectives: a lower bound and an upper bound.

Lower bound. The lower bound indicates the fundamental limitation: $d_{\mathcal{P}}(F^\star, F_{sw})$ cannot be arbitrarily reduced. Firstly, a minimal $d_{\mathcal{P}}(F^\star, F_{sw})$ is intrinsically tied to the weak model performance $d_{\mathcal{P}}(F^\star, F_w)$. To improve the strong model, the weak model becomes critical—that is, $d_{\mathcal{P}}(F^\star, F_w)$ should be as small as possible. It underscores the importance of ***carefully selecting the weak model*** (Burns et al., 2023; Yang et al., 2025). Secondly, the performance improvement of strong model over the weak model cannot exceed $\mathcal{O}\left(\sqrt{d_{\mathcal{P}}(F_w, F_{sw})}\right)$. In W2SG, while the student-supervisor disagreement $d_{\mathcal{P}}(F_w, F_{sw})$ is minimized in Equation (6), we anticipate $\mathcal{O}\left(\sqrt{d_{\mathcal{P}}(F_w, F_{sw})}\right)$ to remain relatively small. However, a paradox arises: achieving a smaller $d_{\mathcal{P}}(F^\star, F_{sw})$ necessitates a larger $d_{\mathcal{P}}(F_w, F_{sw})$. This implies that ***the performance improvement of W2SG is probably constrained by its own optimization objective***.

Upper bound. The upper bound provides a theoretical guarantee for W2SG by ensuring that $d_{\mathcal{P}}(F^\star, F_{sw})$ remains bounded and does not grow arbitrarily large. Firstly, effective W2SG requires choosing a weak model that produces supervision signal closely aligned with the true score, i.e., achieving a small $d_{\mathcal{P}}(F^\star, F_w)$. To this end, employing a stronger weak model is crucial to obtain a tighter upper bound of $d_{\mathcal{P}}(F^\star, F_{sw})$. Secondly, the worst-case performance of the strong model is constrained by the sum of $d_{\mathcal{P}}(F^\star, F_w)$ and $\mathcal{O}\left(\sqrt{d_{\mathcal{P}}(F_w, F_{sw})}\right)$. By appropriately selecting the weak model and determining the minimizer of Equation (6), both $d_{\mathcal{P}}(F^\star, F_w)$ and $\mathcal{O}\left(\sqrt{d_{\mathcal{P}}(F_w, F_{sw})}\right)$ can be kept small, ensuring the practicality of the strong model.

4.2 Calibration Error Bound

In this section, we further explore W2SG through the lens of calibration (Kumar et al., 2019), which requires that the prediction confidence should match the actual outcome. We first state the definition of Marginal Calibration Error (MCE) (Kumar et al., 2019), which is an extended version of Expected Calibration Error (ECE) (Guo et al., 2017) designed for multi-class classification. In particular, we use an ℓ_1 version of MCE, with the weight constant $\frac{1}{k}$ omitted.

Definition 4.1 (Marginal Calibration Error (Kumar et al., 2019))**.** *Let $x \in \mathcal{X}$, ground truth $y = [y_1, \cdots, y_k]^T \in \{0,1\}^k$ where $\sum_{i=1}^k y_i = 1$, and a model $f : \mathcal{X} \to \mathcal{Y}$. Define the marginal calibration error of f as:*

$$MCE(f) = \sum_{i=1}^{k} \mathbb{E}_x \left| [f(x)]_i - \mathbb{P}[y_i = 1 | [f(x)]_i] \right| . \tag{8}$$

It measures the difference between model confidence and actual outcome, and $MCE(f) \in [0, 2]$. For binary classification, *MCE* is twice the *ECE*. We shed light on upper and lower bounds of calibration of the strong model.

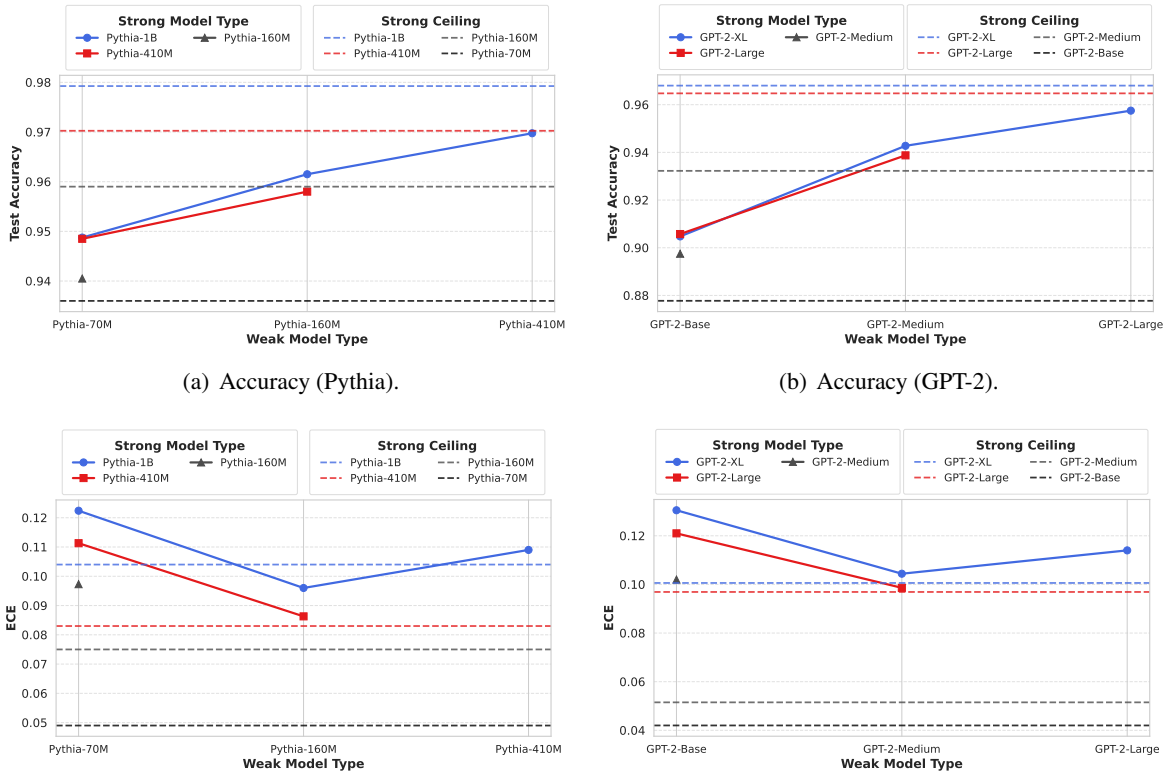
Theorem 4.2 (Proved in Appendix B.3)**.** *Let $MCE(\cdot)$ be the marginal calibration error in Definition 4.1. Then there holds*

$$\begin{aligned} &|MCE(F_{sw}) - MCE(F_w)| \\ &\qquad \leq 2 \cdot \sqrt{1 - \exp\left(-d_{\mathcal{P}}(F_w, F_{sw})\right)}. \end{aligned} \tag{9}$$

Theorem 4.2 demonstrates that the calibration error of F_{sw} is influenced by two key factors: (1) the calibration error of F_w, and (2) the teacher-student disagreement, as characterized by the optimization result in Equation (6). This theoretical result yields two insights. First, to achieve a strong model with acceptable calibration, the weak teacher should also exhibit acceptable calibration. Otherwise, the strong model will inherit a non-trivial calibration error from the weak teacher as $d_{\mathcal{P}}(F_w, F_{sw})$ goes to zero. Second, closely imitating the weak supervisor minimizes $d_{\mathcal{P}}(F_w, F_{sw})$, causing the calibration errors of the strong and weak models to converge. Taking them together, to ensure W2SG with reasonable calibration and prevent a poorly-calibrated F_{sw}, it is crucial to ***avoid using a poorly-calibrated weak model with an overfitted strong model***. Additionally, since models with larger capacity may exhibit higher calibration errors (Guo et al., 2017), a potential trade-off may exist between the weak model's calibration error and the teacher-student disagreement. In other words, $MCE(F_w)$ and $\sqrt{1 - \exp\left(-d_{\mathcal{P}}(F_w, F_{sw})\right)}$ may not be minimized simultaneously, posing a challenge in selecting the weak model and designing an effective optimization strategy to achieve better calibration in the strong model.

4.3 Experimental Validation in Language Models

In this section, we use language models to verify our theoretical results in W2SG.



(a) Accuracy (Pythia).

(b) Accuracy (GPT-2).

(c) ECE (Pythia).

(d) ECE (GPT-2).

Figure 1: Accuracy and calibration results for Pythia and GPT-2 series. (**a**) Test accuracies of Pythia series. (**b**) Test accuracies of GPT-2 series. Each curve demonstrates the variation in accuracy of W2SG as strong models are supervised by weak models of varying capabilities. "Strong Ceiling" corresponds to models fine-tuned using ground truth data. (**c**) Expected calibration errors of Pythia series. Each curve depicts the change in ECE as strong models are supervised by different weak teachers. (**d**) Expected calibration errors of GPT-2 series.

4.3.1 Experimental Setting

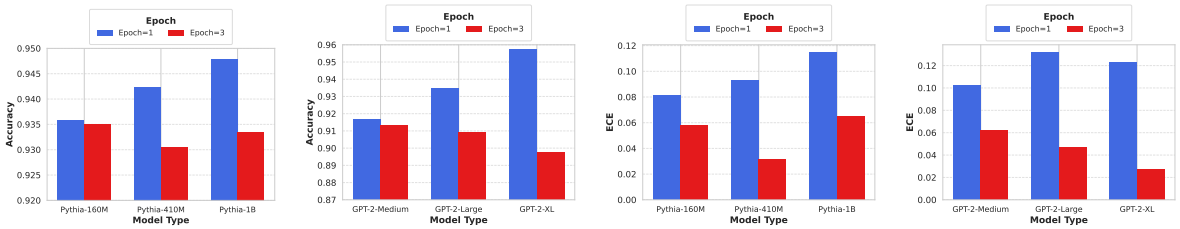
Dataset. We define the alignment objective as enabling a weak model to guide a strong model in achieving harmlessness. To this end, we employ CAI-Harmless (Bai et al., 2022b), which is a widely adopted single-turn harmless dataset for reward modeling task. Each sample is structured as $(x; y_c, y_r)$, where x denotes the prompt, and y_c and y_r represent the human-preferred and human-rejected completions, respectively. The dataset is randomly split into three 4K-sample subsets: one for fine-tuning both weak and strong base models, another for weak supervision via weak model predictions, and the last for testing and evaluation. The dataset is randomly divided into three distinct subsets: (1) 4K ground truth samples for fine-tuning both weak and strong base language models; (2) A held-out set of 4K samples, where labels are predicted by the weak model and used to provide weak supervision for training the strong model; (3) The remaining 4K samples, reserved for testing and evaluating the performance of all models.

Model. To explore weak-to-strong generalization, we utilize GPT-2 series (Radford et al., 2019) (including GPT-2-Base, GPT-2-Medium, GPT-2-Large, and GPT-2-XL) and Pythia series (Biderman et al., 2023) (including Pythia-70M, Pythia-160M, Pythia-410M and Pythia-1B). For each model, we append a linear projection head to facilitate logit predictions for each completion pair $\tilde{x} = (x; y_c, y_r)$. Consequently, the task can be framed as a binary classification problem, where the model F predicts the soft label as

$$F(\tilde{x}) = \text{Sigmoid}(F(y_c) - F(y_r)).$$

Training. The models are trained via KL divergence loss. Training details are in Appendix C.1.

Metric. To evaluate whether a model F can effectively distinguish between chosen and rejected completions (y_c and y_r) for a given prompt x, we aim for F to assign a higher score to the chosen completion compared to the rejected one. Specifically, this requires $F(y_c) - F(y_r) > 0$ for each completion pair $\tilde{x} = (x; y_c, y_r)$, which implies



(a) Accuracy (Pythia). (b) Accuracy (GPT-2). (c) ECE (Pythia). (d) ECE (GPT-2).

Figure 2: Ablation study for the Pythia and GPT-2 series. (**a**)-(**b**) Test accuracies of Pythia and GPT-2. The accuracies of Pythia-70M and GPT-2-Base fine-tuned on ground truth data is 92.45% and 90.95%, respectively. (**c**)-(**d**) ECE of Pythia and GPT-2. The ECE of Pythia-70M and GPT-2-Base fine-tuned on ground truth data is 0.049 and 0.042, respectively.

$F(\tilde{x}) > 0.5$. Accordingly, the test accuracy of a model F is reported as the fraction of predictions that satisfy $F(\tilde{x}) > 0.5$.

4.3.2 Results and Analysis

The generalization and calibration performance for Pythia and GPT-2 series are shown in Figure 1. To further investigate how the optimization result $d_{\mathcal{P}}(F_w, F_{sw})$ affect generalization and calibration, we increase the number of epochs to train a strong model that more closely imitates the weak model. The corresponding results are in Figure 2.

Main results. Figure 1(a) and Figure 1(b) demonstrate that, for the same strong model, the generalization of W2SG increases when supervised by a weak model of greater capacity. This experimental result is consistent with Theorem 4.1. Interestingly, in Figure 1(c) and Figure 1(d), we observe that for the same weak model as the teacher, a stronger student model with higher capacity tends to exhibit larger ECE. This aligns with previous findings on the negative correlation between model capacity and calibration (Guo et al., 2017).

Ablation study. We extend the training epochs and design a series of teacher-student pairs with increasing model capacities. In particular, we employ Pythia-70M as the weak teacher to supervise Pythia-160M, 410M, and 1B. We also utilize GPT-2-Base as the weak teacher for supervising GPT-2-Medium, Large, and XL. Figure 2 illustrates that as we increase the number of epochs to reduce $d_{\mathcal{P}}(F_w, F_{sw})$, *there is a simultaneous decline in both the accuracy and calibration error of other strong models*. Taking the Pythia series as an example, Figure 1(a) and Figure 1(c) demonstrate that Pythia-70M achieves the lowest accuracy and best ECE performance among the Pythia models. While Theorem 4.2 indicates that reducing $d_{\mathcal{P}}(F_w, F_{sw})$ causes the accuracy and calibration results of strong models to converge toward those of the weak model, our experiments show that increasing the number of epochs leads to reduced accuracy and ECE for Pythia-160M, 410M, and 1B. In other words, the accuracy and ECE of strong models approach those of the weak model, consistent with Theorem 4.1 and Theorem 4.2. And this trend is also observed in the GPT-2 series.

Overfitting blocks W2SG. As the number of epochs increases, *the accuracy of GPT-2-XL drops even below that of GPT-2-Base* (90.95%). This is attributed to the strong expressive power of GPT-2-XL, which leads to overfitting to the weak supervision provided by GPT-2-Base. Note that the upper bounds derived in Theorem 4.1 and Theorem 4.2 do not guarantee that the strong model will outperform the weak model in terms of both generalization performance and calibration properties. The underlying intuition is that if a strong model overfits to the weak supervision, it may closely mimic the weak model's generalization and calibration behavior. Consequently, the strong model could end up performing on par with or potentially even worse than the weak model.

5 Results Beyond Squared Loss

In regression problems under some assumptions, Charikar et al. (2024) proves that the strong model's error is smaller than the weak model's, with the gap at least the strong model's error on the weak labels. This observation naturally raises the following question: *Can their proof be extended from squared loss to output distribution divergence?* In this section, we show how to theoretically bridge the gap between squared loss and KL divergence within the overall proof framework established in Charikar et al. (2024). To begin with, we restate an assumption used in previous study.

Assumption 5.1 (Convexity Assumption (Charikar et al., 2024))**.** *The strong model learns fine-tuning tasks from a convex set $\mathcal{F}_s$.*

To satisfy this assumption, $\mathcal{F}_s$ can be the class of all linear functions (Charikar et al., 2024). It is validated by practice: a popular way to fine-tune a pre-trained model on task-specific data is by tuning the last linear layer of the model (Howard and Ruder, 2018; Kumar et al., 2022).

5.1 Upper Bound (Realizability)

Firstly, we consider the case where $\exists f_s \in \mathcal{F}_s$ such that $F_s = F^\star$ (also called "Realizability" (Charikar et al., 2024)). It means we can find a f_s such that $f_s \circ h_s = f^\star \circ h^\star$. This assumption implicitly indicates the strong power of pre-training. It requires that the representation h_s has learned extremely enough information during pre-training, which is reasonable in modern large language models pre-trained on very large corpus (Touvron et al., 2023; Achiam et al., 2023). We state our result in the realizable setting, which corresponds to Theorem 1 in Charikar et al. (2024).

Theorem 5.1 (Proved in Appendix B.4)**.** *Given $F^\star$, F_w and F_{sw} defined above. Consider $\mathcal{F}_s$ that satisfies Assumption 5.1. Consider W2SG using reverse KL divergence loss:*

$$f_{sw} = \mathit{argmin}_{f \in \mathcal{F}_s} \, d_{\mathcal{P}}(f \circ h_s, f_w \circ h_w).$$

Assume that $\exists f_s \in \mathcal{F}_s$ such that $F_s = F^\star$. Then

$$d_{\mathcal{P}}(F^\star, F_{sw}) \leq d_{\mathcal{P}}(F^\star, F_w) - d_{\mathcal{P}}(F_{sw}, F_w). \tag{10}$$

Remark. *The corresponding theorem and proof in the case of forward KL divergence loss is provided in Corollary B.1 from Appendix B.5, under an additional assumption.*

In contrast to the symmetric squared loss studied in prior work (Charikar et al., 2024), the emergence of the reverse KL divergence is inherently tied to the asymmetric properties of the KL divergence. Although extending previous work to both forward and reverse KL divergences presents significant technical challenges, our results demonstrate the theoretical guarantees of W2SG in these settings. In Inequality (10), the left-hand side represents the error of the weakly-supervised strong model on the true data. On the right-hand side, the first term denotes the true error of the weak model, while the second term captures the disagreement between the strong and weak models, which also serves as the minimization objective in W2SG. This inequality indicates that the weakly-supervised strong model improves upon the weak model by at least the magnitude of their disagreement, $d_{\mathcal{P}}(F_{sw}, F_w)$. To reduce the error of F_{sw}, Theorem 5.1 aligns with Theorem 4.1, highlighting the importance of selecting an effective weak model and the inherent limitations of the optimization objective in W2SG.

5.2 Upper Bound (Non-Realizability)

We relax the "realizability" condition and draw n i.i.d. samples. We provide the result in the "unrealizable" setting, where the condition $F_s = F^\star$ may not be satisfied for any $f_s \in \mathcal{F}_s$. It corresponds to Theorem 2 in Charikar et al. (2024).

Theorem 5.2 (Proved in Appendix B.6)**.** *Given $F^\star$, F_w and F_{sw} defined above. Consider $\mathcal{F}_s$ that satisfies Assumption 5.1. Consider weak-to-strong generalization using reverse KL:*

$$f_{sw} = \mathit{argmin}_{f \in \mathcal{F}_s} \, d_{\mathcal{P}}(f \circ h_s, f_w \circ h_w),$$
$$\hat{f}_{sw} = \mathit{argmin}_{f \in \mathcal{F}_s} \, \hat{d}_{\mathcal{P}}(f \circ h_s, f_w \circ h_w),$$

Denote $d_{\mathcal{P}}(F^\star, F_s) = \varepsilon$. With probability at least $1 - \delta$ over the draw of n i.i.d. samples, there holds

$$d_{\mathcal{P}}(F^\star, \hat{F}_{sw}) \leq d_{\mathcal{P}}(F^\star, F_w) - d_{\mathcal{P}}(\hat{F}_{sw}, F_w)$$
$$+ \mathcal{O}(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

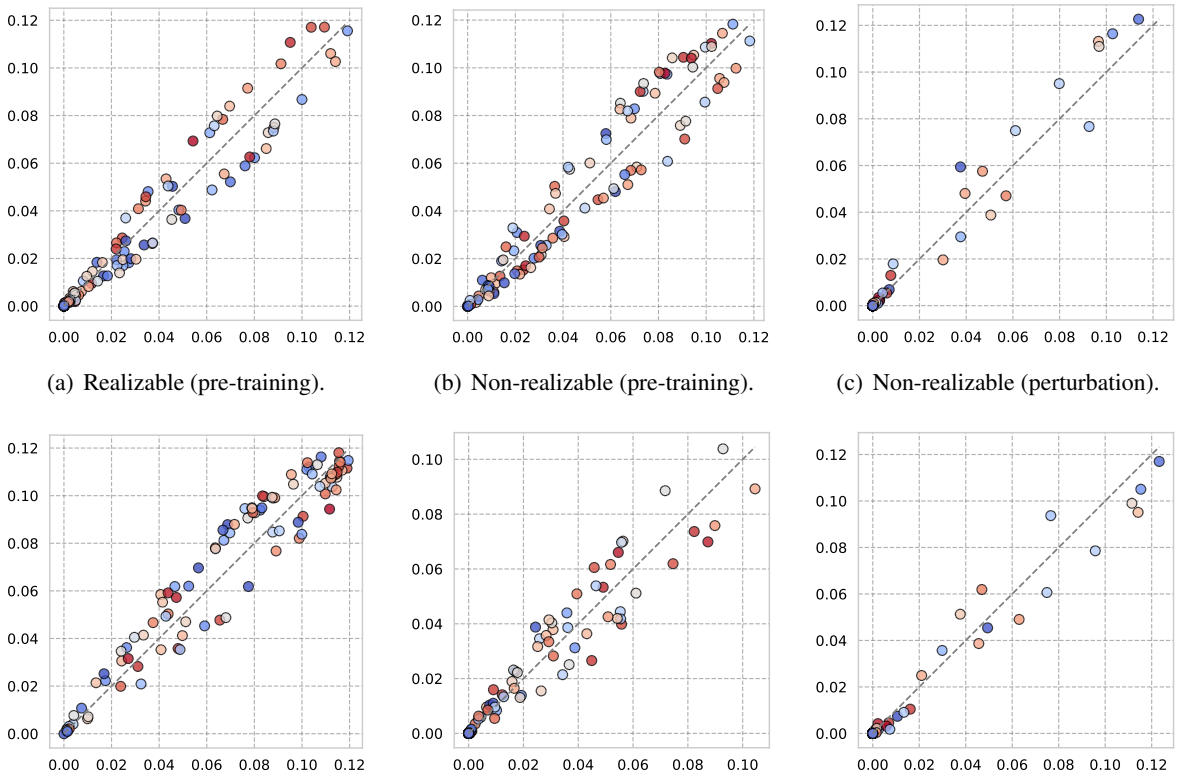
where $\mathcal{C}_{\mathcal{F}_s}$ is a constant capturing the complexity of the function class $\mathcal{F}_s$, and the asymptotic notation is with respect to $\varepsilon \to 0, n \to \infty$.

Remark. *The extension to forward KL divergence loss is provided in Corollary B.2 from Appendix B.7, under an additional assumption.*

Compared to Inequality (10), this bound introduces three another error terms: $\mathcal{O}(\sqrt{\varepsilon})$ arises due to the non-realizability assumption, and diminishes as the strong ceiling model F_s becomes more expressive. The remaining two error terms arise from the strong model $\hat{F}_{sw}$ being trained on a finite weakly-labeled sample. They also asymptotically approach zero as the sample size increases.

5.3 Synthetic Experiments

In this section, we conduct experiments on synthetically generated data to validate the theoretical results in Section 5. We extend the synthetic experiments from Charikar et al. (2024) using exactly the same setting except for the loss



(a) Realizable (pre-training). (b) Non-realizable (pre-training). (c) Non-realizable (perturbation).

(d) Realizable (pre-training). (e) Non-realizable (pre-training). (f) Non-realizable (perturbation).

Figure 3: Synthetic experiments following Charikar et al. (2024) using reverse KL divergence loss (**a-c**) and forward KL divergence loss (**d-f**). Each point corresponds to a task and the gray dotted line represents $y = x$. $h^\star$ is a 16-layer MLP. (**a,d**) Realizable (pre-training): $h_s = h^\star$, and h_w is a 2-layer MLP obtained by pre-training. (**b,e**) Non-realizable (pre-training): h_s is an 8-layer MLP, and h_w is a 2-layer MLP. Both h_s and h_w are obtained by pre-training. (**c,f**) Non-realizable (perturbation): Both h_s and h_w are obtained by directly perturbing the weights in $h^\star$: $h_s = h^\star + \mathcal{N}(0, 0.01)$, and $h_w = h^\star + \mathcal{N}(0, 9)$.

function defined in Equation (4). In particular, the data follows a zero-mean normal distribution. The weak, strong and ground truth representations (h_w, h_s and $h^\star$) are implemented as randomly initialized MLPs. One 'Realizable' and two 'Non-realizable' settings are considered. The full details are in Appendix C.2. To validate Theorem 5.1-5.2 and visualize the trend clearly, we plot $d_{\mathcal{P}}(F^\star, F_w) - d_{\mathcal{P}}(F^\star, F_{sw})$ on the x-axis versus $d_{\mathcal{P}}(F_{sw}, F_w)$ on the y-axis. The results are presented in Figure 3(a)-(c). We also examine forward KL divergence loss in Figure 3(d)-(f).

For the reverse KL divergence loss, the points in cluster around the line $y = x$. This suggests that $d_{\mathcal{P}}(F^\star, F_w) - d_{\mathcal{P}}(F^\star, F_{sw}) \approx d_{\mathcal{P}}(F_{sw}, F_w)$, which aligns with previous results (Charikar et al., 2024) for squared loss. It is also consistent with our theoretical results in Section 5, suggesting that the improvement over the weak teacher can be quantified by the disagreement between strong and weak models. For the forward KL divergence loss, the observed trend closely mirrors that of reverse KL. The dots also are around the line $y = x$, suggesting the same relationship for the forward KL.

6 Conclusion

This paper provides a theoretical analysis of W2SG. In the classification setting, we establish upper and lower bounds for both generalization and calibration errors of the strong model, revealing that the primary limitations arise from the weak model and the optimization objective. These bounds emphasize two critical insights: (1) the weak model must demonstrate strong generalization and calibration performance, and (2) the strong model should avoid excessive training to prevent overfitting on weak supervision. In the regression setting, we extend previous work to output distribution divergence loss, proving that a strong model can outperform its weak teacher by at least their disagreement under certain assumptions. Overall, we hope this work deepens the understanding of W2SG and inspires future research on its foundations.

Limitations

While our work offers theoretical insights and empirical validation in W2SG, several limitations warrant discussion. First, although the theoretical framework in Section 4 provides broad conceptual understanding, the assumptions in Section 5 may not fully align with real-world LLM deployment scenarios. This challenge, however, is common across most theoretical analyses of W2SG. Despite this limitation, our findings establish a foundational framework for future research to refine W2SG theory in LLMs. Second, our empirical evaluation is limited to two established alignment-focused binary classification tasks using relatively small-scale models. While the results validate our theory, further investigation is needed to assess the generalizability of our approach to more diverse datasets and larger model architectures. Addressing this in future studies will be critical for validating the broader applicability of our theory.

Broader Impact and Ethics Statement

This work on weak-to-strong generalization aims to improve the alignment of superhuman models with human values. While our theoretical and empirical insights highlight the potential of this approach, we acknowledge the risks of propagating biases or errors from the weak model to the strong model. To address these concerns, we emphasize the importance of ensuring the weak model's generalization and calibration, as well as carefully balancing the strong model's optimization to avoid over-reliance on weak supervision. We encourage rigorous testing, transparency, and ongoing monitoring in real-world applications to ensure the safe and ethical deployment of such systems. Our work contributes to the broader effort of aligning advanced AI with human values, but its implementation must prioritize fairness, accountability, and safety.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aakriti Agrawal, Mucong Ding, Zora Che, Chenghao Deng, Anirudh Satheesh, John Langford, and Furong Huang. 2024. Ensemw2s: Can an ensemble of llms be leveraged to obtain a stronger llm? *arXiv preprint arXiv:2410.04571*.

Gholamali Aminian, Mahed Abroshan, Mohammad Mahdi Khalili, Laura Toni, and Miguel Rodrigues. 2022. An information-theoretical approach to semi-supervised learning under covariate-shift. In *International Conference on Artificial Intelligence and Statistics*, pages 7433–7449.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Peter L Bartlett and Shahar Mendelson. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. 2022. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10925–10934.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. 2020. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, and 1 others. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. 2024. Quantifying the gain in weak-to-strong generalization. *Advances in neural information processing systems*.

Qi Chen, Changjian Shui, and Mario Marchand. 2021. Generalization bounds for meta-learning: An information-theoretic analysis. *Advances in Neural Information Processing Systems*, 34:25878–25890.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A close look into the calibration of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367.

Lele Cheng, Xiangzeng Zhou, Liming Zhao, Dangwei Li, Hong Shang, Yun Zheng, Pan Pan, and Yinghui Xu. 2020. Weakly supervised learning with side information for noisy labeled images. In *The European Conference on Computer Vision*, pages 306–321.

P Chu and D Messerschmitt. 1982. A frequency weighted itakura-saito spectral distance measure. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(4):545–560.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.

Inderjit S. Dhillon. 2007. Learning with bregman divergences. https://www.cs.utexas.edu/~inderjit/Talks/bregtut.pdf.

Monroe D Donsker and SR Srinivasa Varadhan. 1983. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212.

Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. 2009. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

Maurice Günder, Nico Piatkowski, and Christian Bauckhage. 2022. Full kullback-leibler-divergence loss for hyperparameter-free label distribution learning. *arXiv preprint arXiv:2209.02055*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2021. An overview of uncertainty calibration for text classification and the role of distillation. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 289–306.

Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. 2024. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*.

Yue Guo and Yi Yang. 2024. Improving weak-to-strong generalization with reliability-aware alignment. *arXiv preprint arXiv:2406.19032*.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. 2023. Generalization bounds: Perspectives from information theory and pac-bayes. *arXiv preprint arXiv:2309.04381*.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Ehsan Imani and Martha White. 2018. Improving regression performance with distributional losses. In *International conference on machine learning*, pages 2157–2166.

Fumitada Itakura. 1968. Analysis synthesis telephony based on the maximum likelihood method. *Reports of the 6-th Int. Cong. Acoust*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

HyunJin Kim, Xiaoyuan Yi, Jing Yao, Jianxun Lian, Muhua Huang, Shitong Duan, JinYeong Bak, and Xing Xie. 2024. The road to artificial superintelligence: A comprehensive survey of superalignment. *arXiv preprint arXiv:2412.16468*.

Yoshiaki Kitazawa. 2025. Bounds on l_p errors in density ratio estimation via f-divergence loss functions. In *The Thirteenth International Conference on Learning Representations*.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. 2024. Theoretical analysis of weak-to-strong generalization. *Advances in neural information processing systems*.

Michel Ledoux and Michel Talagrand. 2013. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. 2021. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119.

Shaojie Li, Bowei Zhu, and Yong Liu. 2024. Algorithmic stability unleashed: Generalization bounds with unbounded losses. In *Forty-first International Conference on Machine Learning*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Lydia T Liu, Max Simchowitz, and Moritz Hardt. 2019. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060.

Yuejiang Liu and Alexandre Alahi. 2024. Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts. *arXiv preprint arXiv:2402.15505*.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. 2019. Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31(9):3732–3740.

Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878.

Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong. 2019. Conditional teacher-student learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6445–6449.

Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. 2019. Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. *arXiv preprint arXiv:1906.10606*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2901 – 2907.

OpenAI. 2024. Introducing superalignment.

Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. 2017. Training object class detectors with click supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6374–6383.

Martin Pawelczyk, Lillian Sun, Zhenting Qi, Aounon Kumar, and Himabindu Lakkaraju. 2024. Generalizing trust: Weak-to-strong trustworthiness in language models. *arXiv preprint arXiv:2501.00418*.

Ankit Pensia, Varun Jog, and Po-Ling Loh. 2018. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory*, pages 546–550.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Dani Prasetyawan and Nakamoto Takamichi. 2020. Sensory evaluation of odor approximation using nmf with kullback-leibler divergence and itakura-saito divergence in mass spectrum space. *Journal of The Electrochemical Society*, 167(16):167520.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. 2022. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4036–4054.

Daniel Russo and James Zou. 2016. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240.

Jitao Sang, Yuhang Wang, Jing Zhang, Yanxu Zhu, Chao Kong, Junhong Ye, Shuyu Wei, and Jinlin Xiao. 2024. Improving weak-to-strong generalization with scalable oversight and ensemble learning. *arXiv preprint arXiv:2402.00667*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. 2018. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*.

Seamus Somerstep, Felipe Maia Polo, Moulinath Banerjee, Ya'acov Ritov, Mikhail Yurochkin, and Yuekai Sun. 2024. A statistical framework for weak-to-strong generalization. *arXiv preprint arXiv:2405.16236*.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153.

Huayi Tang and Yong Liu. 2023. Information-theoretic generalization bounds for transductive learning and its applications. *arXiv preprint arXiv:2311.04561*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in nlp: A study on the impact of method & data scarcity. *arXiv preprint arXiv:2210.15452*.

Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440.

Ziqiao Wang and Yongyi Mao. 2023a. Information-theoretic analysis of unsupervised domain adaptation. In *International Conference on Learning Representations*.

Ziqiao Wang and Yongyi Mao. 2023b. Tighter information-theoretic generalization bounds from supersamples. In *International Conference on Machine Learning*.

David X Wu and Anant Sahai. 2025. Provable weak-to-strong generalization via benign overfitting. In *International Conference on Learning Representations*.

Aolin Xu and Maxim Raginsky. 2017. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30.

Wenkai Yang, Shiqi Shen, Guangyao Shen, Wei Yao, Yong Liu, Zhi Gong, Yankai Lin, and Ji-Rong Wen. 2025. Super (ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. In *International Conference on Learning Representations*.

Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. 2021. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. In *Advances in Neural Information Processing Systems*.

Yuqing Yang, Yan Ma, and Pengfei Liu. 2024. Weak-to-strong reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8350–8367.

Ruimeng Ye, Yang Xiao, and Bo Hui. 2024. Weak-to-strong generalization beyond accuracy: a pilot study in safety, toxicity, and legal reasoning. *arXiv preprint arXiv:2410.12621*.

Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795.

Wenhong Zhu, Zhiwei He, Xiaofeng Wang, Pengfei Liu, and Rui Wang. 2025. Weak-to-strong preference optimization: Stealing reward from weak aligned model. In *International Conference on Learning Representations*.

Contents

1 Introduction **1**

2 Related Work **2**

3 Preliminaries **2**
- 3.1 Classification and Regression . . . 2
- 3.2 Weak-to-strong Generalization . . . 3

4 Universal Results in W2SG **3**
- 4.1 Generalization Error Bounds . . . 3
- 4.2 Calibration Error Bound . . . 4
- 4.3 Experimental Validation in Language Models . . . 4
  - 4.3.1 Experimental Setting . . . 5
  - 4.3.2 Results and Analysis . . . 6

5 Results Beyond Squared Loss **6**
- 5.1 Upper Bound (Realizability) . . . 7
- 5.2 Upper Bound (Non-Realizability) . . . 7
- 5.3 Synthetic Experiments . . . 7

6 Conclusion **8**

A Further Related Work **14**

B Main Proof **15**
- B.1 Proof of Theorem 4.1 . . . 15
- B.2 Extension of Theorem 4.1 . . . 17
- B.3 Proof of Theorem 4.2 . . . 18
- B.4 Proof of Theorem 5.1 . . . 19
- B.5 Extension of Theorem 5.1 . . . 20
- B.6 Proof of Theorem 5.2 . . . 25
- B.7 Extension of Theorem 5.2 . . . 30

C Further Details and Results of Experiments **33**
- C.1 Training Details of Experiments in Language Models . . . 33
- C.2 Weak-to-Strong Training Procedure in Synthetic Experiments . . . 33

Appendix

A Further Related Work

Teacher-student learning paradigm. The student-teacher training paradigm (Meseguer-Brocal et al., 2019; Meng et al., 2019), which involves first training a teacher model and then using its outputs (e.g., pseudo-labels or soft targets) to guide the training of a student model, has become a cornerstone in various machine learning domains. This approach is particularly prominent in knowledge distillation (Hinton, 2015; Beyer et al., 2022), semi-supervised learning (Tarvainen and Valpola, 2017) and domain adaptation (Shu et al., 2018). It can also be used in other fields like curriculum learning (Matiisen et al., 2019) and continual learning (Lee et al., 2021). However, most prior work assumes that the teacher is either more capable or at least comparable to the student in terms of model capacity or performance. In contrast, weak-to-strong generalization (Burns et al., 2023) explores a less studied setting where the student model is significantly more capable than the teacher, which is different from the traditional assumptions of the student-teacher framework. By theoretically investigating this setting, we aim to uncover novel insights into the capabilities and limitations of weak-to-strong generalization. We hope that our theoretical investigation into it will serve as a catalyst for advancements not only in the domain of super-alignment but also in the broader landscape of teacher-student learning paradigm.

Weakly-supervised learning. Weakly-supervised learning has emerged as a powerful paradigm to address the challenges of limited labeled data by leveraging weak supervision (Ratner et al., 2020). Such weak supervision may be incomplete (i.e., only a small subset of labels are given), inexact (i.e., only coarse-grained labels are given) and inaccurate (i.e., the given labels are noisy) (Zhou, 2018). This problem setting is also closely related to label noise (Song et al., 2022) and semi-supervised learning (Van Engelen and Hoos, 2020). To address the problem of weakly-supervised learning, practical trials leverage these various forms of weak supervision, such as utilizing noisy labels (Cheng et al., 2020), coarse-grained labels (Oquab et al., 2015), and incomplete annotations (Papadopoulos et al., 2017). And most of them improving model performance within the limitations of weak supervision. In contrast, weak-to-strong generalization explores a distinct yet related direction: it investigates how a strong model, when trained on weak supervision, can not only correct the errors of the weak supervisor but also generalize to instances where the weak supervisor is uncertain or incorrect (Burns et al., 2023; Yang et al., 2025). We hope that our theoretical exploration of weak-to-strong generalization can inspire not only the field of super-alignment but also research in weakly-supervised learning.

Calibration. Calibration is an important concept about uncertainty estimation and confidence (Guo et al., 2017; Kuleshov et al., 2018; Kumar et al., 2019; Mehrtash et al., 2020) in machine learning. There are several kinds of definition for calibration. For instance, taking expectation conditioned on the data distribution (Kull et al., 2019; Kumar et al., 2019; Roelofs et al., 2022) (Also, see Definition 3 in (Pleiss et al., 2017) and Equation (2) in (Liu et al., 2019) in the fairness literature), and (2) taking expectation conditioned on the probability score (Naeini et al., 2015; Guo et al., 2017). Researchers also investigate the calibration in natural language processing (Desai and Durrett, 2020; Guo et al., 2021; Ulmer et al., 2022; Chen et al., 2023). In recent years, the calibration of large language models has garnered significant attention (Zhu et al., 2023; Tian et al., 2023; Liang et al., 2023), with a thorough survey provided in (Geng et al., 2024). However, to the best of our knowledge, while confidence issues in weak-to-strong generalization have been investigated in (Burns et al., 2023), the role of calibration has not been sufficiently investigated. In this paper, we theoretically demonstrate how strong model's calibration is affected in W2SG. And we believe that calibration warrants further in-depth investigation in this field.

Information-theoretic analysis. Information-theoretic analysis is commonly employed to bound the expected generalization error in supervised learning (Russo and Zou, 2016; Xu and Raginsky, 2017), with subsequent studies providing sharper bounds (Bu et al., 2020; Wang and Mao, 2023b). These bounds have been used to characterize the generalization ability of stochastic gradient-based optimization algorithms (Pensia et al., 2018). Furthermore, this theoretical framework has been extended to diverse settings, including meta-learning (Chen et al., 2021), semi-supervised learning (Aminian et al., 2022), transductive learning (Tang and Liu, 2023), and domain adaptation (Wang and Mao, 2023a). For a

comprehensive overview of these developments, we refer readers to the recent monograph by Hellström et al. (2023). Nonetheless, despite its extensive application across various domains, the information-theoretic analysis of super-alignment (OpenAI, 2024), particularly in the context of weak-to-strong generalization, remains largely underexplored. In this paper, we use KL divergence to analyze weak-to-strong generalization, which is not considered in previous work. KL divergence is an information-theoretic measure between two probability distributions in information theory (Cover, 1999). And how to extend it to other information-theoretic measures remains an open question and warrants further exploration in future work.

B Main Proof

B.1 Proof of Theorem 4.1

We first state some preliminaries for the proof.

Lemma B.1 (Donsker and Varadhan's variational formula (Donsker and Varadhan, 1983))**.** *Let Q, P be probability measures on $\mathcal{X}$, for any bounded measurable function $f : \mathcal{X} \to \mathbb{R}$, we have*

$$\mathrm{D}_{\mathrm{KL}}(Q\|P) = \sup_f \mathbb{E}_{x\sim Q}[f(x)] - \log \mathbb{E}_{x\sim P}[\exp f(x)].$$

Lemma B.2 (Hoeffding's lemma)**.** *Let $X \in \mathbb{R}$ such that $a \le X \le b$. Then, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Definition B.1 (Subgaussian random variable)**.** *A random variable $X \in \mathbb{R}$ is σ-subgaussian if for any ρ,*

$$\log \mathbb{E}\exp(\rho(X - \mathbb{E}X)) \le \rho^2\sigma^2/2.$$

Notation of probability distribution for the model output. We define the corresponding probability distributions for prediction of F_{sw} and F_w. Recall that for $F_w, F_{sw} : \mathcal{X} \to \mathcal{Y}$ and $x \in \mathcal{X}$:

$$d_{\mathcal{P}}(F_w, F_{sw}) = \mathbb{E}_x\left[\sum_{j=1}^{k}[F_w(x)]_j \log\frac{[F_w(x)]_j}{[F_{sw}(x)]_j}\right] = \mathbb{E}_x\left[\mathrm{D}_{\mathrm{KL}}(F_w(x), F_{sw}(x))\right],$$

where D_{KL} is the discrete version of KL divergence. $\forall x \in \mathcal{X}$, we know that $\sum_{j=1}^{k}[F_w(x)]_j = 1$. Therefore, given the class space $C_k = \{1, \cdots, k\}$, we define a probability distribution $\mathcal{P}_w(x)$ with the probability density function p_w, where $j \in C_k$ and

$$p_w(j) = [F_w(x)]_j. \tag{11}$$

Using this method, we also define the probability distribution $\mathcal{P}_{sw}(x)$ for $F_w(x)$.

Now we start the proof.

Proof. For better readability, we divide the proof into several steps.

The first step. Given the probability distributions $\mathcal{P}_w(x)$ and $\mathcal{P}_{sw}(x)$ above, the first step is motivated by Lemma A.2 from (Wang and Mao, 2023a). For any $x \in \mathcal{X}$, $j \in C_k$, $g : C_k \to \mathbb{R}$ and assume that g is σ-subgaussian (we will specify σ later). Let $f = t \cdot g$ for any $t \in \mathbb{R}$. We have

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left(F_w(x)\|F_{sw}(x)\right)\\
=&\mathrm{D}_{\mathrm{KL}}(\mathcal{P}_w(x)\|\mathcal{P}_{sw}(x))\\
=&\sup_t \mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[t\cdot g(j')\right] - \log\mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[\exp\left(t\cdot g(j)\right)] && \text{(Lemma B.1)}\\
=&\sup_t \mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[tg\left(j'\right)\right] - \log\mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}\left[\exp t\left(g(j) - \mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)] + \mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)]\right)\right]\\
=&\sup_t \mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[tg\left(j'\right)\right] - \mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[tg(j)] - \log\mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}\left[\exp t\left(g(j) - \mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)]\right)\right]\\
\ge&\sup_t t\left(\mathbb{E}_{j'\sim\mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_{j\sim\mathcal{P}_{sw}(x)}[g(j)]\right) - t^2\sigma^2/2. && \text{(Subgaussianity)}
\end{aligned}$$

The second step. The second step is associating the above result with $d_{\mathcal{P}}(F_w, F_{sw})$. In particular, by taking expectations of x on both sides of the above inequality, we obtain

$$d_{\mathcal{P}}(F_w, F_{sw}) = \mathbb{E}_x \mathrm{D}_{\mathrm{KL}}\left(F_w(x) \| F_{sw}(x)\right) \geq \sup_t \underbrace{t\left(\mathbb{E}_x \mathbb{E}_{j' \sim \mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_x \mathbb{E}_{j \sim \mathcal{P}_{sw}(x)}[g(j)]\right) - t^2\sigma^2/2}_{\phi(t)}.$$

Note that $\phi(t)$ is a quadratic function of t. Therefore, by AM–GM inequality, we find the maximum of this quadratic function:

$$\phi(t) \leq \frac{1}{2\sigma^2}\left(\mathbb{E}_x \mathbb{E}_{j' \sim \mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_x \mathbb{E}_{j \sim \mathcal{P}_{sw}(x)}[g(j)]\right)^2 = \sup_t \phi(t) \leq d_{\mathcal{P}}(F_w, F_{sw}).$$

Subsequently, there holds

$$\left|\mathbb{E}_x \mathbb{E}_{j' \sim \mathcal{P}_w(x)}\left[g\left(j'\right)\right] - \mathbb{E}_x \mathbb{E}_{j \sim \mathcal{P}_{sw}(x)}[g(j)]\right| \leq \sqrt{2\sigma^2 d_{\mathcal{P}}(F_w, F_{sw})}. \tag{12}$$

The third step. The third step is constructing g to associate the above result with $d_{\mathcal{P}}(F^\star, F_{sw})$ and $d_{\mathcal{P}}\left(F^\star, F_w\right)$. Specifically, given a probability distribution $\mathcal{P}_g$ with the density function p_g, we define function $g : C_k \to (0, 1]$ associated with $\mathcal{P}_g$:

$$g(j) \triangleq \frac{[F^\star(x)]_j}{p_g(j)} \log \frac{[F^\star(x)]_j}{p_g(j)}, \quad \text{for } j \in C_k.$$

We have

$$\begin{aligned}
\mathbb{E}_x \mathbb{E}_{j \sim \mathcal{P}_g}\left[g(j)\right] &= \mathbb{E}_x \mathbb{E}_{j \sim \mathcal{P}_g}\left[\frac{[F^\star(x)]_j}{p_g(j)} \log \frac{[F^\star(x)]_j}{p_g(j)}\right] \\
&= \mathbb{E}_x\left[\sum_{j \in C_k} p_g(j) \cdot \frac{[F^\star(x)]_j}{p_g(j)} \cdot \log \frac{[F^\star(x)]_j}{p_g(j)}\right] \\
&= \mathbb{E}_x\left[\sum_{j \in C_k} [F^\star(x)]_j \cdot \log \frac{[F^\star(x)]_j}{p_g(j)}\right]
\end{aligned}$$

Recall the definition of $\mathcal{P}_{sw}$ and $\mathcal{P}_w$ in (11), we replace $\mathcal{P}_g$ with $\mathcal{P}_{sw}$ and $\mathcal{P}_w$ in the above equation:

$$\mathbb{E}_x \mathbb{E}_{j' \sim \mathcal{P}_{sw}}\left[g\left(j'\right)\right] = \mathbb{E}_x\left[\sum_{j=1} [F^\star(x)]_j \log \frac{[F^\star(x)]_j}{[F_{sw}(x)]_j}\right] = d_{\mathcal{P}}(F^\star, F_{sw}),$$

$$\mathbb{E}_x \mathbb{E}_{j \sim \mathcal{P}_w}[g(j)] = \mathbb{E}_x\left[\sum_{j=1} [F^\star(x)]_j \log \frac{[F^\star(x)]_j}{[F_w(x)]_j}\right] = d_{\mathcal{P}}(F^\star, F_w).$$

Substitute the above into (12):

$$|d_{\mathcal{P}}(F^\star, F_{sw}) - d_{\mathcal{P}}(F^\star, F_w)| \leq \sqrt{2\sigma^2 d_{\mathcal{P}}(F_w, F_{sw})}. \tag{13}$$

The final step. Finally, we obtain the subgaussian factor R of function g by using the fact that g is bounded. For simplicity, we use Hoeffding's Lemma (Lemma B.2) to obtain the subgaussian factor R. However, it can be more precisely determined using advanced techniques in learning theory literature (for instance, see Remark 2.14 in (Li et al., 2024), where $\alpha = 2$ recovers the subgaussian setting).

Recall that the output domain $\mathcal{Y} \subseteq \mathbb{R}^k$, where $\forall y = (y_1, \cdots, y_k)^T \in \mathcal{Y}$, there holds $\sum_{i=1}^k y_i = 1$ and $0 < y_i \leq 1$. In other words, $\exists \gamma > 0$ such that $0 < \gamma \leq y_i \leq 1$. It means that $g(j) \in [-\frac{1}{\gamma}\log\frac{1}{\gamma}, \frac{1}{\gamma}\log\frac{1}{\gamma}]$. According to Lemma B.2, $\forall \lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda(g(j) - \mathbb{E}[g(j)])}\right] \leq \exp\left(\frac{\lambda^2\left(\frac{1}{\gamma}\log\frac{1}{\gamma}\right)^2}{2}\right).$$

In other words, $g(j)$ is σ-subgaussian, where $\sigma = \frac{1}{\gamma}\log\frac{1}{\gamma}$. Substitute it into Equation (13) and we obtain:

$$|d_{\mathcal{P}}(F^\star, F_{sw}) - d_{\mathcal{P}}(F^\star, F_w)| \leq C_1\sqrt{d_{\mathcal{P}}(F_w, F_{sw})},$$

where the constant $C_1 = \frac{\sqrt{2}}{\gamma}\log\frac{1}{\gamma}$. The proof is complete. □

B.2 Extension of Theorem 4.1

In this section, we extend Theorem 4.1 to output distribution divergence in regression.

Proof. Denote $\mathrm{d_{KL}}(f\|g) = \mathrm{D_{KL}}(\mathcal{P}_f\|\mathcal{P}_g) = \int_{\mathcal{X}} f(x)\log\frac{f(x)}{g(x)}dx$. Let $(\mathcal{X}, \mathcal{F}, \mathcal{P}_{sw}), (\mathcal{X}, \mathcal{F}, \mathcal{P}_w)$ be two probability spaces. Denoting F_{sw} and F_w the densities of the measures. Therefore,

$$\int_{\mathcal{X}} F_{sw}(x)dx = \int_{\mathcal{X}} F_w(x)dx = 1.$$

Let $x \in \mathcal{X}$, $g : \mathcal{X} \to \mathbb{R}$ and assume that g is R-subgaussian (we will specify R later). Let $f = t \cdot g$ for any $t \in \mathbb{R}$. By Lemma B.1, we have

$$\begin{aligned}
\mathrm{d_{KL}}\left(F_w\|F_{sw}\right) &= \mathrm{D_{KL}}(\mathcal{P}_w\|\mathcal{P}_{sw}) \\
&= \sup_t \mathbb{E}_{x'\sim\mathcal{P}_w}\left[t\cdot g\left(x'\right)\right] - \log\mathbb{E}_{x\sim\mathcal{P}_{sw}}[\exp\left(t\cdot g(x)\right)] \\
&= \sup_t \mathbb{E}_{x'\sim\mathcal{P}_w}\left[tg\left(x'\right)\right] - \log\mathbb{E}_{x\sim\mathcal{P}_{sw}}\left[\exp t\left(g(x) - \mathbb{E}_{x\sim\mathcal{P}_{sw}}[g(x)] + \mathbb{E}_{x\sim\mathcal{P}_{sw}}[g(x)]\right)\right] \\
&= \sup_t \mathbb{E}_{x'\sim\mathcal{P}_w}\left[tg\left(x'\right)\right] - \mathbb{E}_{x\sim\mathcal{P}_{sw}}[tg(x)] - \log\mathbb{E}_{x\sim\mathcal{P}_{sw}}\left[\exp t\left(g(x) - \mathbb{E}_{x\sim\mathcal{P}_{sw}}[g(x)]\right)\right] \\
&\geq \sup_t \underbrace{t\left(\mathbb{E}_{x'\sim\mathcal{P}_w}\left[g\left(x'\right)\right] - \mathbb{E}_{x\sim\mathcal{P}_{sw}}[g(x)]\right) - t^2R^2/2}_{\phi(t)}. \qquad \text{(Subgaussianity)}
\end{aligned}$$

Let

$$\phi(t) = t\left(\mathbb{E}_{x'\sim\mathcal{P}_w}\left[g\left(x'\right)\right] - \mathbb{E}_{x\sim\mathcal{P}_{sw}}[g(x)]\right) - t^2R^2/2,$$

which is a quadratic function of t. Therefore,

$$\phi(t) \leq \frac{1}{2R^2}\left(\mathbb{E}_{x'\sim\mathcal{P}_{sw}}\left[g\left(x'\right)\right] - \mathbb{E}_{x\sim\mathcal{P}_w}[g(x)]\right)^2 = \sup_t \phi(t) \leq \mathrm{D_{KL}}(\mathcal{P}_w\|\mathcal{P}_{sw}) = \mathrm{d_{KL}}\left(F_w\|F_{sw}\right).$$

So we have

$$\left|\mathbb{E}_{x'\sim\mathcal{P}_{sw}}\left[g\left(x'\right)\right] - \mathbb{E}_{x\sim\mathcal{P}_w}[g(x)]\right| \leq \sqrt{2R^2\mathrm{d_{KL}}\left(F_w\|F_{sw}\right)}. \tag{14}$$

Given a probability space $(\mathcal{X}, \mathcal{F}, \mathcal{P}_g)$ with the density function p_g. Define

$$g(x) \triangleq \frac{F^\star(x)}{p_g(x)}\log\frac{F^\star(x)}{p_g(x)}.$$

So there holds

$$\mathbb{E}_{x\sim\mathcal{P}_g}\left[g(x)\right] = \int_{\mathcal{X}} F^\star(x)\cdot\log\frac{F^\star(x)}{p_g(x)}dx.$$

Replace $\mathcal{P}_g$ with $\mathcal{P}_{sw}$ and $\mathcal{P}_w$:

$$\mathbb{E}_{x'\sim\mathcal{P}_{sw}}\left[g\left(x'\right)\right] = \int_{\mathcal{X}} F^\star(x)\log\frac{F^\star(x)}{F_{sw}(x)}dx = d_{\mathcal{P}}(F^\star, F_{sw}),$$

$$\mathbb{E}_{x\sim\mathcal{P}_w}[g(x)] = \int_{\mathcal{X}} F^\star(x)\log\frac{F^\star(x)}{F_w(x)}dx = d_{\mathcal{P}}(F^\star, F_w).$$

Substitute them back into (14) and we obtain:

$$|d_{\mathcal{P}}(F^\star, F_{sw}) - d_{\mathcal{P}}(F^\star, F_w)| \leq \sqrt{2R^2\mathrm{d_{KL}}\left(F_w\|F_{sw}\right)}. \tag{15}$$

Finally, recall that the output domain $\mathcal{Y} = \{y \in \mathbb{R} | 0 < y \leq 1\}$. In other words, $\exists \gamma > 0$ such that $\mathcal{Y} = \{y \in \mathbb{R} | 0 < \gamma \leq y \leq 1\}$. It means that $g(x) \in [-\frac{1}{\gamma}\log\frac{1}{\gamma}, \frac{1}{\gamma}\log\frac{1}{\gamma}]$. According to Lemma B.2, $\forall \lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda(g(x)-\mathbb{E}[g(x)])}\right] \leq \exp\left(\frac{\lambda^2\left(\frac{1}{\gamma}\log\frac{1}{\gamma}\right)^2}{2}\right).$$

In other words, $g(x)$ is R-subgaussian, where $R = \frac{1}{\gamma}\log\frac{1}{\gamma}$. Substitute it into Equation (15) and we obtain:

$$|d_{\mathcal{P}}(F^\star, F_{sw}) - d_{\mathcal{P}}(F^\star, F_w)| \leq \sqrt{C_1 d_{\mathcal{P}}(F_w, F_{sw})},$$

where the constant $C_1 = 2\left(\frac{1}{\gamma}\log\frac{1}{\gamma}\right)^2$. □

B.3 Proof of Theorem 4.2

Total variation distance is introduced for our proof.

Definition B.2 (Total Variation Distance)**.** *Given two probability distributions P and Q, the Total Variation (TV) distance between P and Q is*

$$\mathrm{D}_{\mathrm{TV}}(P\|Q) = \frac{1}{2}\int_{x\in\mathcal{X}} |P(x) - Q(x)|\, dx.$$

Note that $\mathrm{D}_{\mathrm{TV}}(P\|Q) \in [0,1]$. Also, $\mathrm{D}_{\mathrm{TV}}(P\|Q) = 0$ if and only if P and Q coincides, and $\mathrm{D}_{\mathrm{TV}}(P\|Q) = 1$ if and only if P and Q are disjoint.

Let the calibrated bayes score function $F^b : \mathcal{X} \to \mathcal{Y}$ that satisfies $\forall i \in \{1, \cdots, k\}$, $[F^b(x)]_i = \mathbb{P}(Y_i = 1|X = x)$, where $Y = [Y_1, \cdots, Y_k]^T \in \{0,1\}^k$ and $\|Y\|_1 = 1$. Now we start our proof.

Proof. Consider the definition of MCE in Equation (8). Notice that

$$\begin{aligned}
\mathit{MCE}(F) &= \sum_{i=1}^{k} \mathbb{E}_X \left|[F(X)]_i - \mathbb{P}[Y_i = 1|[F(X)]_i]\right| \\
&= \mathbb{E}_X\left[\sum_{i=1}^{k} \left|[F(X)]_i - \mathbb{P}[Y_i = 1|[F(X)]_i]\right|\right] \\
&= \mathbb{E}_X \left\|F(X) - F^b(X)\right\|_1,
\end{aligned}$$

and $\mathit{MCE}(F) \in [0,2]$.

So there holds

$$\begin{aligned}
\mathit{MCE}(F_w) - \mathit{MCE}(F_{sw}) &= \mathbb{E}_X \left\|F_w(X) - F^b(X)\right\|_1 - \mathbb{E}_X \left\|F_{sw}(X) - F^b(X)\right\|_1 \\
&\leq \mathbb{E}_X \left\|F_w(X) - F_{sw}(X)\right\|_1 && \text{(Triangle inequality)} \\
&= 2\cdot\mathbb{E}_X \mathrm{D}_{\mathrm{TV}}(F_w(X), F_{sw}(X)) \\
&\leq 2\cdot\mathbb{E}_X \sqrt{1 - \exp\left(-\mathrm{D}_{\mathrm{KL}}(F_w(X), F_{sw}(X))\right)} \\
& && \text{(Bretagnolle–Huber inequality)} \\
&\leq 2\cdot\sqrt{1 - \exp\left(-\mathbb{E}_X \mathrm{D}_{\mathrm{KL}}(F_w(X), F_{sw}(X))\right)} && \text{(Jensen's inequality)} \\
&= 2\cdot\sqrt{1 - \exp\left(-d_{\mathcal{P}}(F_w, F_{sw})\right)}. && \text{(Definition of } d_{\mathcal{P}} \text{ for KL divergence loss)}
\end{aligned}$$

Likewise,

$$\begin{aligned}
\mathit{MCE}(F_{sw}) - \mathit{MCE}(F_w) &= \mathbb{E}_X \left\|F_{sw}(X) - F^b(X)\right\|_1 - \mathbb{E}_X \left\|F_w(X) - F^b(X)\right\|_1 \\
&\leq \mathbb{E}_X \left\|F_{sw}(X) - F_w(X)\right\|_1 && \text{(Triangle inequality)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}_X \left\| F_w(X) - F_{sw}(X) \right\|_1 && \text{(Symmetry)} \\
&= 2 \cdot \mathbb{E}_X \mathrm{D}_{\mathrm{TV}}(F_w(X), F_{sw}(X)) \\
&\leq 2 \cdot \sqrt{1 - \exp\left(-d_{\mathcal{P}}(F_w, F_{sw})\right)}. && \text{(Using the derivation above)}
\end{aligned}
$$

Combining the above, we have

$$
\begin{aligned}
\mathit{MCE}(F_{sw}) - \mathit{MCE}(F_w) &\leq 2 \cdot \sqrt{1 - \exp\left(-d_{\mathcal{P}}(F_w, F_{sw})\right)}, \\
\mathit{MCE}(F_w) - \mathit{MCE}(F_{sw}) &\leq 2 \cdot \sqrt{1 - \exp\left(-d_{\mathcal{P}}(F_w, F_{sw})\right)}.
\end{aligned}
$$

The proof is complete.

□

B.4 Proof of Theorem 5.1

We first restate a lemma for our proof. Recall that the strong model learns from a linear function class $\mathcal{F} : \mathbb{R}^{d_s} \to \mathbb{R}$ of fine-tuning tasks. Recall also that we denote the strong model representation map by $h_s : \mathbb{R}^d \to \mathbb{R}^{d_s}$. Let $V_s = \{f \circ h_s : f \in \mathcal{F}\}$ be the set of all tasks in $\mathcal{F}$ composed with the strong model representation. We first observe that V_s is also a convex set.

Lemma B.3 (Charikar et al. (2024)). *V_s is a convex set.*

Proof. Fix $f, g \in \mathcal{F}$, and consider $f \circ h_s, g \circ h_s \in V_s$. Fix any $\lambda \in [0, 1]$. Since $\mathcal{F}$ is the linear function class so that it is a convex set, there exists $p \in \mathcal{F}$ such that for all $y \in \mathbb{R}^{d_s}$, $p(y) = \lambda f(y) + (1 - \lambda) g(y)$. Now, fix any $x \in \mathbb{R}^d$. Then, we have that

$$
\lambda(f \circ h_s)(x) + (1 - \lambda)(g \circ h_s)(x) = \lambda f(h_s(x)) + (1 - \lambda) g(h_s(x)) = p(h_s(x)) = (p \circ h_s)(x),
$$

and hence $\lambda(f \circ h_s) + (1 - \lambda)(g \circ h_s) = p \circ h_s \in V_s$. □

Now we start the proof.

Proof. For any $f, g \in \mathcal{X} \to \mathcal{Y}$, denote $\mathrm{d}_{\mathrm{KL}}(f \| g) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx$.

Given any $g \in V_s$, observe that

$$
\begin{aligned}
\mathrm{d}_{\mathrm{KL}}(g \| F_w) &= \int_{\mathcal{X}} g(x) \log \frac{g(x)}{F_w(x)} dx \\
&= \int_{\mathcal{X}} g(x) \log \left(\frac{g(x)}{F_{sw}(x)} \cdot \frac{F_{sw}(x)}{F_w(x)} \right) dx \\
&= \mathrm{d}_{\mathrm{KL}}(g \| F_{sw}) + \mathrm{d}_{\mathrm{KL}}(F_{sw} \| F_w) - \mathrm{d}_{\mathrm{KL}}(F_{sw} \| F_w) + \int_{\mathcal{X}} g(x) \log \frac{F_{sw}(x)}{F_w(x)} dx \\
&= \mathrm{d}_{\mathrm{KL}}(g \| F_{sw}) + \mathrm{d}_{\mathrm{KL}}(F_{sw} \| F_w) + \underbrace{\int_{\mathcal{X}} (g(x) - F_{sw}(x)) \log \frac{F_{sw}(x)}{F_w(x)} dx}_{Q_1}. \qquad (16)
\end{aligned}
$$

Now our goal is to judge whether $Q_1 \geq 0$.

Recall that

$$
f_{sw} = \operatorname{argmin}_f \mathrm{d}_{\mathrm{KL}}(f \circ h_s \| F_w).
$$

In other words, F_{sw} is the *projection* of F_w onto the convex set V_s. Therefore:

$$
\mathrm{d}_{\mathrm{KL}}(g \| F_w) \geq \mathrm{d}_{\mathrm{KL}}(F_{sw} \| F_w).
$$

Therefore,

$$
\mathrm{d}_{\mathrm{KL}}(g \| F_{sw}) + Q_1 \geq 0. \qquad (17)
$$

Now, fix $t \in (0, 1)$, and consider the function $g = F_{sw} + t(F^\star - F_{sw})$.

$$
\begin{aligned}
\mathrm{d}_{\mathrm{KL}}(g\|F_{sw}) &= \int_{\mathcal{X}} g(x)\log\frac{g(x)}{F_{sw}(x)} \\
&= -\int_{\mathcal{X}} g(x)\log\frac{F_{sw}(x)}{g(x)}dx \\
&= -\int_{\mathcal{X}} g(x)\log\left[1-\frac{t(F^\star(x)-F_{sw}(x))}{g(x)}\right]dx \\
&= \int_{\mathcal{X}} g(x)\left[\frac{t(F^\star(x)-F_{sw}(x))}{g(x)}+\mathcal{O}(t^2)\right]dx \\
&= \int_{\mathcal{X}} \left[t(F^\star(x)-F_{sw}(x))+\mathcal{O}(t^2)\right]dx && \text{(Taylor expansion)} \\
&= \mathcal{O}(t^2). && (\textstyle\int_{\mathcal{X}} F_{sw}(x)dx = \int_{\mathcal{X}} F^\star(x)dx = 1)
\end{aligned}
$$

While

$$
\begin{aligned}
Q_1 &= \int_{\mathcal{X}} (g(x)-F_{sw}(x))\log\frac{F_{sw}(x)}{F_w(x)}dx \\
&= t\cdot\int_{\mathcal{X}} (F^\star(x)-F_{sw}(x))\log\frac{F_{sw}(x)}{F_w(x)}dx \\
&= \mathcal{O}(t).
\end{aligned}
$$

Recall Equation (17) that

$$
\underbrace{\mathrm{d}_{\mathrm{KL}}(F_{sw}\|g)}_{\mathcal{O}(t^2)} + \underbrace{Q_1}_{\mathcal{O}(t)} \geq 0,
$$

which means $Q_1 \geq 0$. So we have

$$
\int_{\mathcal{X}} (F^\star(x)-F_{sw}(x))\log\frac{F_{sw}(x)}{F_w(x)}dx \geq 0.
$$

Let $g = F^\star$ in Equation (16) and we can prove the result $d_{\mathcal{P}}(F^\star, F_{sw}) \leq d_{\mathcal{P}}(F^\star, F_w) - d_{\mathcal{P}}(F_{sw}, F_w)$. □

B.5 Extension of Theorem 5.1

We first introduce some definitions.

Definition B.3 (Itakura–Saito Divergence (Itakura, 1968; Févotte et al., 2009; Prasetyawan and Takamichi, 2020))**.** *Given two probability distributions P and Q, the Itakura–Saito divergence between them is defined as*

$$
\mathrm{D}_{\mathrm{IS}}(P\|Q) = \int_{\mathcal{X}} \left(\frac{P(x)}{Q(x)} - \log\frac{P(x)}{Q(x)} - 1\right)dx.
$$

Similar to the KL divergence, the Itakura–Saito divergence is also a Bregman divergence (Dhillon, 2007).

Definition B.4 (Weighted Itakura–Saito Divergence (Chu and Messerschmitt, 1982))**.** *Given two probability distributions P and Q, the weighted Itakura–Saito divergence between them is defined as*

$$
\mathrm{D}_{\mathrm{WIS}}(P\|Q) = \int_{\mathcal{X}} w(x)\left(\frac{P(x)}{Q(x)} - \log\frac{P(x)}{Q(x)} - 1\right)dx,
$$

where $w(x)$ is the weight function.

We also define the inner product of functions

$$
\langle f, g\rangle \triangleq \int_{\mathcal{X}} f(x)g(x)dx.
$$

Now we present the theoretical extension of Theorem 5.1 to forward KL divergence loss in W2SG.

Corollary B.1. *Given $F^\star$, F_w and F_{sw} defined above. Let $d_{\mathcal{P}}$ be the output distribution divergence and consider W2SG in Equation (6) using forward KL divergence loss. Assume that $\exists f_s \in \mathcal{F}_s$ such that $F_s = F^\star$. Consider $\mathcal{F}_s$ that satisfies Assumption 5.1. If the weighted Itakura–Saito divergence $\mathrm{D}_{\mathrm{WIS}}(F^\star \| F_{sw}) \leq 0$ with the weight function $w = F_{sw} - F_w$, then:*

$$d_{\mathcal{P}}\left(F_{sw}, F^\star\right) \leq d_{\mathcal{P}}\left(F_w, F^\star\right) - d_{\mathcal{P}}\left(F_w, F_{sw}\right). \tag{18}$$

Corollary B.1 shows that Charikar et al. (2024) can be extended to our setting if we introduce another assumption, which comes from the technical challenges of theoretically analyzing the non-linear nature of KL divergence. Denote $F^+ = \frac{F^\star}{F_{sw}} - 1 - \log \frac{F^\star}{F_{sw}}$, then the assumption $\mathrm{D}_{\mathrm{WIS}}(F^\star \| F_{sw}) \leq 0$ is equivalent to $\langle F_{sw} - F_w, F^+ \rangle \leq 0$. Note that $\forall x \in \mathcal{X}$, $F_+(x) \geq 0$ always holds, and a very small or large value of $\frac{F^\star(x)}{F_{sw}(x)}$ generally contributes to a large $F_+(x)$. To make $\langle F_{sw} - F_w, F^+ \rangle \leq 0$ more likely to hold, we expect $F_{sw}(x) \leq F_w(x)$ if $F^\star(x)$ is small or large. In general, since this condition cannot be guaranteed to hold universally, the inequality in Equation (10) may fail to hold. This reveals a key discrepancy between the square function (as considered in Charikar et al. (2024)) and the KL divergence (in this work) within the W2SG framework—a phenomenon that will be empirically validated through our experiments.

Proof sketch of Charikar et al. (2024). For the proof technique, Charikar et al. (2024) constructs a function within a convex set. By exploiting the property of projection and square function, they demonstrate that $\mathcal{O}(t) + \mathcal{O}(t^2)$ is non-negative as $t \to 0^+$. Consequently, the first-order term must be non-negative, which proves the result.

Proof sketch of ours. Extending the proof framework from Theorem 1 in (Charikar et al., 2024) presents several challenges. First, due to the properties of KL divergence, the constructed function does not lie within the convex set. To address this issue, we employ a first-order Taylor expansion and introduce a remainder term. Secondly, because of the remainder, we derive that $\mathcal{O}(t) + \mathcal{O}(t) + \mathcal{O}(t^2)$ is non-negative. Consequently, we must assume that one of the first-order terms is non-positive to ensure that the other first-order term is non-negative, which allows us to prove the result. However, if the first-order term is positive, the second first-order term might also remain non-negative.

Now we start our proof of Corollary B.1. Some Taylor expansion claims used in the proof (Claim B.1, Claim B.2 and Claim B.3) are provided at the end of the proof.

Proof. For any $f, g \in \mathcal{X} \to \mathcal{Y}$, denote $\mathrm{d}_{\mathrm{KL}}(f \| g) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx$.

Given any $g \in V_s$, observe that

$$\begin{aligned}
\mathrm{d}_{\mathrm{KL}}(F_w \| g) &= \int_{\mathcal{X}} F_w(x) \log \frac{F_w(x)}{g(x)} dx \\
&= \int_{\mathcal{X}} F_w(x) \log \left(\frac{F_w(x)}{F_{sw}(x)} \cdot \frac{F_{sw}(x)}{g(x)} \right) dx \\
&= \int_{\mathcal{X}} F_w(x) \log \frac{F_w(x)}{F_{sw}(x)} dx + \int_{\mathcal{X}} F_w(x) \log \frac{F_{sw}(x)}{g(x)} dx \\
&= \mathrm{d}_{\mathrm{KL}}(F_w \| F_{sw}) + \mathrm{d}_{\mathrm{KL}}(F_{sw} \| g) - \mathrm{d}_{\mathrm{KL}}(F_{sw} \| g) + \int_{\mathcal{X}} F_w(x) \log \frac{F_{sw}(x)}{g(x)} dx \\
&= \mathrm{d}_{\mathrm{KL}}(F_w \| F_{sw}) + \mathrm{d}_{\mathrm{KL}}(F_{sw} \| g) + \underbrace{\int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \log \frac{F_{sw}(x)}{g(x)} dx}_{Q_1}.
\end{aligned} \tag{19}$$

Now our goal is to judge whether $Q_1 \geq 0$.

Recall that

$$f_{sw} = \mathrm{argmin}_f \, \mathrm{d}_{\mathrm{KL}}(F_w \| f \circ h_s).$$

In other words, F_{sw} is the *projection* of F_w onto the convex set V_s. Therefore:

$$\mathrm{d}_{\mathrm{KL}}(F_w \| g) \geq \mathrm{d}_{\mathrm{KL}}(F_w \| F_{sw}).$$

And hence

$$
\begin{aligned}
&\mathrm{d}_{\mathrm{KL}}(F_w \| g) - \mathrm{d}_{\mathrm{KL}}(F_w \| F_{sw}) \geq 0 \\
\Rightarrow &\int_{\mathcal{X}} F_w(x) \log \frac{F_w(x)}{g(x)} dx - \int_{\mathcal{X}} F_w(x) \log \frac{F_w(x)}{F_{sw}(x)} dx \geq 0 \\
\Rightarrow &\underbrace{\int_{\mathcal{X}} F_w(x) \log \frac{F_{sw}(x)}{g(x)} dx}_{Q_2} \geq 0.
\end{aligned}
$$

Therefore,

$$
Q_2 = \mathrm{d}_{\mathrm{KL}}(F_{sw} \| g) + Q_1 \geq 0. \tag{20}
$$

Now, fix $t \in (0, 1)$, and consider the functions

$$
w(x) = (F_{sw}(x)) \cdot \left(\frac{F^\star(x)}{F_{sw}(x)} \right)^t, \tag{21}
$$

$$
w'(x) = F_{sw}(x) + t \left(F^\star(x) - F_{sw}(x) \right). \tag{22}
$$

It is clear that $w(x) > 0$, $w'(x) > 0$. And according to Claim B.1, we have

$$
w(x) = F_{sw}(x) \cdot \left[1 + t \log \frac{F^\star(x)}{F_{sw}(x)} + \frac{1}{2} t^2 \left(\log \frac{F^\star(x)}{F_{sw}(x)} \right)^2 \left(\frac{F^\star(x)}{F_{sw}(x)} \right)^\xi \right],
$$

where $\xi \in (0, t)$. It means that

$$
w'(x) - w(x) = \underbrace{t \cdot F_{sw}(x) \left(\frac{F^\star(x)}{F_{sw}(x)} - 1 - \log \frac{F^\star(x)}{F_{sw}(x)} \right)}_{\mathcal{O}(t)} - \underbrace{t^2 \cdot \frac{F_{sw}(x)}{2} \left(\log \frac{F^\star(x)}{F_{sw}(x)} \right)^2 \left(\frac{F^\star(x)}{F_{sw}(x)} \right)^\xi}_{\mathcal{O}(t^2)}. \tag{23}
$$

And $\frac{F^\star(x)}{F_{sw}(x)} - 1 - \log \frac{F^\star(x)}{F_{sw}(x)} > 0$. It means that as $t \to 0^+$, $w'(x) - w(x) > 0$ and $w'(x) - w(x) = \mathcal{O}(t)$.

Define

$$
\begin{aligned}
\phi(x) &\triangleq \log \frac{F_{sw}(x)}{w'(x)} - \log \frac{F_{sw}(x)}{w(x)} && (24) \\
&= \log \frac{w(x)}{w'(x)} \\
&= \log \left(1 - \frac{w'(x) - w(x)}{w'(x)} \right) \\
&= \underbrace{- \frac{w'(x) - w(x)}{w'(x)}}_{\mathcal{O}(t)} \underbrace{- \frac{1}{2} \left(\frac{w'(\zeta) - w(\zeta)}{w'(\zeta)} \right)^2}_{\mathcal{O}(t^2)}, && \text{(Claim B.2)}
\end{aligned}
$$

where ζ is between 0 and $\frac{w'(x) - w(x)}{w'(x)}$. As $t \to 0^+$, we know that $\phi(x) \leq 0$ and $\phi(x) = \mathcal{O}(t)$.

Combining (21) with (24) and we have:

$$
\begin{aligned}
\log \frac{F_{sw}(x)}{w(x)} &= t \log \frac{F_{sw}(x)}{F^\star(x)}, \\
\text{and} \quad \log \frac{F_{sw}(x)}{w'(x)} &= t \log \frac{F_{sw}(x)}{F^\star(x)} + \phi(x).
\end{aligned} \tag{25}
$$

Note that $F_{sw} \in V_s$, which is a convex set (Lemma B.3). Also, $F^\star \in V_s$ (Realizability). Therefore, according to the definition of convexity, we have $w' \in V_s$. Hence, substituting w' for g in (20) and consider Equation (25), we get

$$Q_2 = \underbrace{\int_{\mathcal{X}} F_{sw}(x) \log \frac{F_{sw}(x)}{g(x)} dx}_{\mathrm{d_{KL}}(F_{sw}\|g)} + \underbrace{\int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \log \frac{F_{sw}(x)}{g(x)} dx}_{Q_1} \geq 0,$$

$$\Rightarrow \quad \mathrm{d_{KL}}(F_{sw}\|w') + \int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \cdot \phi(x) dx + t \int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \log \frac{F_{sw}(x)}{F^\star(x)} dx \geq 0. \tag{26}$$

Here, we address these three components individually. Our goal is to show that the first term is $\mathcal{O}(t^2)$, while the second term, which is $\mathcal{O}(t)$, is negative. Considering these collectively, the third term, also $\mathcal{O}(t)$, must be positive.

The first term. Denote $u(x) = F^\star(x) - F_{sw}(x)$. Note that

$$\begin{aligned}
\mathrm{d_{KL}}(F_{sw}\|w') &= \int_{\mathcal{X}} (F_{sw}(x)) \log \frac{F_{sw}(x)}{F_{sw}(x) + t \cdot u(x)} dx \\
&= \int_{\mathcal{X}} (F_{sw}(x)) \log \left(1 - \frac{t \cdot u(x)}{F_{sw}(x) + t \cdot u(x)}\right) dx \\
&= -t \int_{\mathcal{X}} \frac{F_{sw}(x) \cdot u(x)}{F_{sw}(x) + t \cdot u(x)} dx - \frac{1}{2} t^2 \int_{\mathcal{X}} \frac{F_{sw}(x) \cdot (u(\zeta'))^2}{(w'(\zeta'))^2} dx \qquad \text{(Claim B.2)} \\
&= -t \int_{\mathcal{X}} \left[u(x) - \frac{t \cdot (u(x))^2}{F_{sw}(x) + t \cdot u(x)}\right] dx - \frac{1}{2} t^2 \int_{\mathcal{X}} \frac{F_{sw}(x) \cdot (u(\zeta'))^2}{(w'(\zeta'))^2} dx \\
&= -t \underbrace{\int_{\mathcal{X}} u(x) dx}_{0} + t^2 \int_{\mathcal{X}} \frac{(u(x))^2}{F_{sw}(x) + t \cdot u(x)} dx - t^2 \int_{\mathcal{X}} \frac{F_{sw}(x) \cdot (u(\zeta'))^2}{2(w'(\zeta'))^2} dx \\
&= t^2 \int_{\mathcal{X}} \left[\frac{(u(x))^2}{F_{sw}(x) + t \cdot u(x)} - \frac{F_{sw}(x) \cdot (u(\zeta'))^2}{2(w'(\zeta'))^2}\right] dx
\end{aligned}$$

where ζ' is between 0 and $\frac{t \cdot u(x)}{F_{sw}(x) + t \cdot u(x)}$. Therefore, taking the limit as $t \to 0^+$, we get that $\mathrm{d_{KL}}(F_{sw}\|w') = \mathcal{O}(t^2)$.

The second term. Recall Equation (23) that

$$w'(x) - w(x) = t \cdot F_{sw}(x) \left(\frac{F^\star(x)}{F_{sw}(x)} - 1 - \log \frac{F^\star(x)}{F_{sw}(x)}\right) + \mathcal{O}(t^2),$$

$$w'(x) = F_{sw}(x) \left(1 + t \left(\frac{F^\star(x)}{F_{sw}(x)} - 1\right)\right).$$

Therefore,

$$\begin{aligned}
\frac{w'(x) - w(x)}{w'(x)} &= \frac{t \left(\frac{F^\star(x)}{F_{sw}(x)} - 1 - \log \frac{F^\star(x)}{F_{sw}(x)}\right) + \mathcal{O}(t^2)}{1 + t \left(\frac{F^\star(x)}{F_{sw}(x)} - 1\right)} \\
&= \left[t \left(\frac{F^\star(x)}{F_{sw}(x)} - 1 - \log \frac{F^\star(x)}{F_{sw}(x)}\right) + \mathcal{O}(t^2)\right] \cdot \left[1 - t \left(\frac{F^\star(x)}{F_{sw}(x)} - 1\right) + \mathcal{O}(t^2)\right] \\
&\qquad \text{(Claim B.3)} \\
&= t \left(\frac{F^\star(x)}{F_{sw}(x)} - 1 - \log \frac{F^\star(x)}{F_{sw}(x)}\right) + \mathcal{O}(t^2).
\end{aligned}$$

In other words, the second term in (26) is

$$
\begin{aligned}
&\int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \cdot \phi(x) dx \\
=& -\int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \cdot \left(\frac{w'(x) - w(x)}{w'(x)} + \mathcal{O}(t^2) \right) dx && \text{(Equation (24))} \\
=& \int_{\mathcal{X}} (F_{sw}(x) - F_w(x)) \cdot \left(\frac{w'(x) - w(x)}{w'(x)} \right) dx + \mathcal{O}(t^2) \\
=& t \int_{\mathcal{X}} (F_{sw}(x) - F_w(x)) \left(\frac{F^\star(x)}{F_{sw}(x)} - 1 - \log \frac{F^\star(x)}{F_{sw}(x)} \right) dx + \mathcal{O}(t^2) \\
=& t \cdot \mathrm{D}_{\mathrm{WIS}}(F^\star \| F_{sw}) + \mathcal{O}(t^2),
\end{aligned}
$$

where the weight function of the weighted Itakura–Saito Divergence is $w = F_{sw} - F_w$. Therefore, as $t \to 0^+$, the second term in (26) is of the order $\mathcal{O}(t)$. If $\mathrm{D}_{\mathrm{WIS}}(F^\star \| F_{sw}) \leq 0$, the second term in (26) will be non-positive (otherwise, it will be positive).

The third term. Taking the limit as $t \to 0^+$ in (26):

$$
\underbrace{\mathrm{d}_{\mathrm{KL}}(F_{sw} \| w')}_{\mathcal{O}(t^2)} + \underbrace{\int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \cdot \phi(x) dx}_{\mathcal{O}(t)} + \underbrace{t \int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \log \frac{F_{sw}(x)}{F^\star(x)} dx}_{\mathcal{O}(t)} \geq 0.
$$

If the middle term is non-positive, the last term should be non-negative:

$$
\int_{\mathcal{X}} (F_w(x) - F_{sw}(x)) \log \frac{F_{sw}(x)}{F^\star(x)} dx \geq 0. \tag{27}
$$

Substituting $F^\star$ for g in (19), and using (27), we obtain the desired result

$$
d_{\mathcal{P}}(F_{sw}, F^\star) \leq d_{\mathcal{P}}(F_w, F^\star) - d_{\mathcal{P}}(F_w, F_{sw}).
$$

Else, if the middle term is positive (i.e,, $\mathrm{D}_{\mathrm{WIS}}(F^\star \| F_{sw}) \leq 0$ is not satisfied), the last term may also be non-negative, which means that $d_{\mathcal{P}}(F_{sw}, F^\star) \leq d_{\mathcal{P}}(F_w, F^\star) - d_{\mathcal{P}}(F_w, F_{sw})$ may also hold.

□

The following tools used in the above proof can be proved by Taylor expansion.

Claim B.1. *For $t, x \in \mathbb{R}_+$, there holds*

$$
x^t = 1 + t \log x + \frac{1}{2} t^2 (\log x)^2 x^\xi,
$$

where $\xi \in (0, t)$.

Claim B.2. *For $x \in (0, 1)$, there holds:*

$$
\log(1 - x) = -x - \frac{1}{2} \zeta^2,
$$

where $\zeta \in (0, x)$.

Claim B.3. *For $x \in (-1, 1)$, there holds:*

$$
\frac{1}{1 + x} = 1 - x + \epsilon^2,
$$

where ϵ is between 0 and x.

B.6 Proof of Theorem 5.2

Proof sketch of Charikar et al. (2024). They apply the proof technique from their Theorem 1 to different variables and obtain several inequalities. Subsequently, leveraging the triangle inequality for the ℓ_2-norm and a uniform convergence argument, they establish the desired result.

Proof sketch of ours. Extending the proof framework from Theorem 2 in (Charikar et al., 2024) is also non-trivial. Specifically, the absence of a triangle inequality for KL divergence necessitates an alternative approach. To address this, we decompose the relevant terms in a manner analogous to the triangle inequality and exhaustively demonstrate that each of the three resulting remainder terms asymptotically converges to zero.

Notations. For a clear presentation, let

$$\begin{aligned}
A &= d_{\mathcal{P}}(F_s, F_{sw}) \\
B &= d_{\mathcal{P}}(F_{sw}, F_w) \\
C &= d_{\mathcal{P}}(F_s, F_w) \\
D &= d_{\mathcal{P}}(F^\star, F_s) = \varepsilon \\
E &= d_{\mathcal{P}}(F^\star, F_{sw}) \\
F &= d_{\mathcal{P}}(F^\star, F_w) \\
G &= d_{\mathcal{P}}(F^\star, \hat{F}_{sw}) \\
H &= d_{\mathcal{P}}(\hat{F}_{sw}, F_{sw}) \\
I &= d_{\mathcal{P}}(\hat{F}_{sw}, F_w).
\end{aligned}$$

Now we start the proof of Theorem 5.2. A uniform convergence result and two claims used in the proof (Lemma B.4, Claim B.4 and Claim B.5) are provided at the end of the proof.

Proof. Non-realizable weak-to-strong generalization where $F^\star \notin V_s$, and we use a finite sample to perform weak-to-strong supervision. Note that by virtue of the range of $f^\star$, f_w and all functions in $\mathcal{F}$ being absolutely bounded, and $d_{\mathcal{P}}$ is also bounded.

Due to $F^\star \notin V_s$, we replace $F^\star$ with F_s in the final step of proof of Theorem 5.1, we obtain

$$C \geq A + B. \tag{28}$$

Notice that

$$E = A + D - \underbrace{\int_{\mathcal{X}} (F^\star(x) - F_s(x)) \log \frac{F_{sw}(x)}{F_s(x)} dx}_{t_1}, \tag{29}$$

$$F = C + D - \underbrace{\int_{\mathcal{X}} (F^\star(x) - F_s(x)) \log \frac{F_w(x)}{F_s(x)} dx}_{t_2}, \tag{30}$$

$$G = E - H - \underbrace{\int_{\mathcal{X}} (\hat{F}_{sw}(x) - F^\star(x)) \log \frac{F_{sw}(x)}{\hat{F}_{sw}(x)} dx}_{t_3}. \tag{31}$$

Combining (28) and (29), we get

$$E \leq C + D - B - t_1. \tag{32}$$

By a uniform convergence argument (Lemma B.5), we have that with probability at least $1 - \delta$ over the draw of $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ that were used to construct $\hat{F}_{sw}$,

$$I \leq B + \underbrace{\mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right)}_{t_4} + \underbrace{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)}_{t_5}. \tag{33}$$

Combining (32) with (33) and we have

$$E \leq C + D - I - t_1 + t_4 + t_5. \tag{34}$$

Combining (30) with (34) and we have

$$E \leq F - I - t_1 + t_2 + t_4 + t_5. \tag{35}$$

Combining (31) with (35) and we have

$$G \leq F - I - H - t_1 + t_2 - t_3 + t_4 + t_5. \tag{36}$$

We replace $F^\star$ with $\hat{F}_{sw}$ in the final step of proof of Corollary B.1 and obtain:

$$I \geq H + B. \tag{37}$$

Combining (37) with (33) and we have

$$0 \leq H \leq t_4 + t_5 = \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{38}$$

Combining (38) with (36) and we have

$$G \leq F - I - t_1 + t_2 - t_3 + t_4 + t_5. \tag{39}$$

While t_4 and t_5 are known in (33), we analyze t_1, t_2 and t_3 one by one.

Deal with t_1. We know that

$$t_1 = \int_{\mathcal{X}} (F^\star(x) - F_s(x)) \log \frac{F_{sw}(x)}{F_s(x)} dx.$$

Using Pinsker's inequality and the fact that $\frac{F_{sw}(x)}{F_s(x)} \leq \frac{1}{\gamma}$, we have

$$|t_1| \leq \frac{1}{\gamma} \int_{\mathcal{X}} |F_s(x) - F^\star(x)| \, dx \leq \frac{1}{\gamma} \sqrt{\frac{1}{2} \mathrm{d}_{\mathrm{KL}}(F_s \| F^\star)} = \frac{1}{\gamma} \sqrt{\frac{1}{2} \varepsilon}. \tag{40}$$

Therefore,

$$|t_1| = \mathcal{O}(\sqrt{\varepsilon}). \tag{41}$$

Deal with t_2. The proof for t_2 is similar for t_1. In particular, replacing F_{sw} with F_w in the above and we can get

$$|t_2| = O(\sqrt{\varepsilon}). \tag{42}$$

Deal with t_3. We know that

$$t_3 = \int_{\mathcal{X}} (\hat{F}_{sw}(x) - F^\star(x)) \log \frac{F_{sw}(x)}{\hat{F}_{sw}(x)} dx.$$

According to Lemma B.5, with probability at least $1 - \delta$ over the draw of $(x_1, y_1), \ldots, (x_n, y_n)$, we have

$$\left| d_{\mathcal{P}}(\hat{F}_{sw}, F_w) - d_{\mathcal{P}}(F_{sw}, F_w) \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{43}$$

According to Claim B.4 and (43), we have

$$\left|\int_{\mathcal{X}} \log \frac{F_{sw}(x)}{\hat{F}_{sw}(x)} dx\right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Since $|\hat{F}_{sw}(x) - F^{\star}(x)|$ is upper bounded, there holds

$$|t_3| = \left|\int_{\mathcal{X}} (\hat{F}_{sw}(x) - F^{\star}(x)) \log \frac{F_{sw}(x)}{\hat{F}_{sw}(x)} dx\right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{44}$$

Therefore, combing (41), (42) and (44), we have

$$|t_1| + |t_2| + |t_3| \leq O(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{45}$$

Finally, combing (33) and (39) with (38) and (45), we get the result:

$$d_{\mathcal{P}}(F^{\star}, \hat{F}_{sw}) \leq d_{\mathcal{P}}(F^{\star}, F_w) - d_{\mathcal{P}}(\hat{F}_{sw}, F_w) + O(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where in the last inequality, we instantiate asymptotics with respect to $\varepsilon \to 0$ and $n \to \infty$.

□

Here are some tools used in the above proof.

Lemma B.4 (Uniform convergence (forward KL loss))**.** *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be an i.i.d. training sample, where each $x_i \sim \mathcal{P}$ and $y_i = F_w(x_i)$ for a target function F_w. For a fixed strong model representation h_s, let*

$$f_{sw} = \mathit{argmin}_{f \in \mathcal{F}_s}\, d_{\mathcal{P}}(F_w, f \circ h_s),$$
$$\hat{f}_{sw} = \mathit{argmin}_{f \in \mathcal{F}_s}\, \hat{d}_{\mathcal{P}}(F_w, f \circ h_s).$$

Assume that the range of F_w and functions in $\mathcal{F}_s$ is absolutely bounded. Then, with probability at least $1 - \delta$ over the draw of $(x_1, y_1), \ldots, (x_n, y_n)$, we have

$$\left|d_{\mathcal{P}}(F_w, \hat{F}_{sw}) - d_{\mathcal{P}}(F_w, F_{sw})\right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where $\mathcal{C}_{\mathcal{F}_s}$ is a constant capturing the complexity of the function class $\mathcal{F}_s$.

Proof. The proof is strongly motivated by lemma 4 in Charikar et al. (2024).

Note that

$$\begin{aligned} d_{\mathcal{P}}(F_w, \hat{F}_{sw}) - d_{\mathcal{P}}(F_w, F_{sw}) = &\underbrace{d_{\mathcal{P}}(F_w, \hat{F}_{sw}) - \hat{d}_{\mathcal{P}}(F_w, \hat{F}_{sw})}_{a} + \\ &\underbrace{\hat{d}_{\mathcal{P}}(F_w, \hat{F}_{sw}) - \hat{d}_{\mathcal{P}}(F_w, F_{sw})}_{b} + \underbrace{\hat{d}_{\mathcal{P}}(F_w, F_{sw}) - d_{\mathcal{P}}(F_w, F_{sw})}_{c}. \end{aligned} \tag{46}$$

By the definition of $\hat{f}_{sw}$, the second term $b \leq 0$ in (46). Therefore,

$$\left|d_{\mathcal{P}}(F_w, \hat{F}_{sw}) - d_{\mathcal{P}}(F_w, F_{sw})\right| \leq |a| + |c|. \tag{47}$$

The terms a and c measure the difference between the empirical risk and true population risk, and can be controlled by a standard uniform convergence argument.

Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \sim \mathcal{P}$ and $y_i = F_w(x_i)$. According to statistical learning theory literature (Bartlett and Mendelson, 2002), it first holds that with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}_s} |\hat{d}_{\mathcal{P}}(F_w, f \circ h_s) - d_{\mathcal{P}}(F_w, f \circ h_s)| \leq O\left(\mathcal{R}_n(l(\mathcal{F}_s))\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where $\mathcal{R}_n(l(\mathcal{F}_s))$ is the *Rademacher complexity* of the loss class of $\mathcal{F}_s$:

$$\mathcal{R}_n(l(\mathcal{F}_s)) = \mathbb{E}_S \mathbb{E}_{\varepsilon_i \sim \{-1,1\}} \sup_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \cdot \ell(f \circ h_s(x_i), y_i).$$

Notice again that the model output space $\mathcal{Y} = \{y \in \mathbb{R} | 0 < \gamma \leq y \leq 1, \gamma > 0\}$. We can then use the assumption that the range of F_w and $\mathcal{F}_s$ is absolutely bounded, which implies that ℓ is both bounded and Lipschitz in both arguments. This allows us to use the contraction principle in Theorem 4.12 from Ledoux and Talagrand (2013) so as to move from the Rademacher complexity of the loss class $l(\mathcal{F}_s)$ to that of $\mathcal{F}_s$ itself, and claim that with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}_s} |\hat{d}_{\mathcal{P}}(F_w, f \circ h_s) - d_{\mathcal{P}}(F_w, f \circ h_s)| \leq O\left(\mathcal{R}_n(\mathcal{F}_s)\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right) \tag{48}$$

Finally, the Rademacher complexity $\mathcal{R}_n(\mathcal{F}_s)$ can be upper bounded by a quantity known as the *worst-case Gaussian complexity* of $\mathcal{F}_s$; in any case, for a majority of parametric function classes $\mathcal{F}_s$, this quantity scales as $\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}$ (Bartlett and Mendelson, 2002), where $\mathcal{C}_{\mathcal{F}_s}$ is a constant capturing the inherent complexity of $\mathcal{F}_s$. Plugging this into (48) and considering $f = \hat{f}_{sw}$ or $f = f_{sw}$ in this inequality, we have

$$\underbrace{\left|\hat{d}_{\mathcal{P}}(F_w, \hat{F}_{sw}) - d_{\mathcal{P}}(F_w, \hat{F}_{sw})\right|}_{|a|} \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

$$\underbrace{\left|\hat{d}_{\mathcal{P}}(F_w, F_{sw}) - d_{\mathcal{P}}(F_w, F_{sw})\right|}_{|c|} \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Finally, substitute it into Equation (47) and we can obtain the desired bound. □

Lemma B.5 (Uniform convergence (reverse KL loss))**.** *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be an i.i.d. training sample, where each $x_i \sim \mathcal{P}$ and $y_i = F_w(x_i)$ for a target function F_w. For a fixed strong model representation h_s, let*

$$f_{sw} = \textit{argmin}_{f \in \mathcal{F}_s}\, d_{\mathcal{P}}(f \circ h_s, F_w),$$
$$\hat{f}_{sw} = \textit{argmin}_{f \in \mathcal{F}_s} \hat{d}_{\mathcal{P}}(f \circ h_s, F_w).$$

Assume that the range of F_w and functions in $\mathcal{F}_s$ is absolutely bounded. Then, with probability at least $1 - \delta$ over the draw of $(x_1, y_1), \ldots, (x_n, y_n)$, we have

$$\left|d_{\mathcal{P}}(\hat{F}_{sw}, F_w) - d_{\mathcal{P}}(F_{sw}, F_w)\right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where $\mathcal{C}_{\mathcal{F}_s}$ is a constant capturing the complexity of the function class $\mathcal{F}_s$.

Proof. Swap the order of the two elements in $d_{\mathcal{P}}(\cdot, \cdot)$ and $\hat{d}_{\mathcal{P}}(\cdot, \cdot)$ in the proof of Lemma B.4 and we can prove the result. □

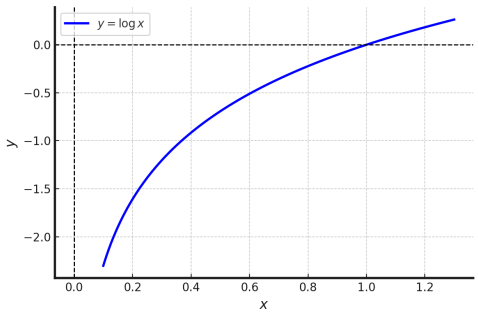


Figure 4: The function $\phi(x) = \log x$.

Claim B.4. *Let $f(x), g(x) \in [\gamma, 1]$ where $\gamma > 0$. If there exists $\xi > 0$ such that $\int_{\mathcal{X}} |f(x) - g(x)|\, dx \leq \xi$, then there holds*

$$\int_{\mathcal{X}} |\log f(x) - \log g(x)|\, dx \leq \frac{1}{\gamma}\xi.$$

Proof. Using the property of the function $\phi(x) = \log x$ (as shown in Figure 4): if $x \in (0, 1]$, then the slope of a line with any two points on the function $\phi(x)$ is bounded.

In particular, we have

$$\begin{aligned}
&\int_{\mathcal{X}} |\log f(x) - \log g(x)|\, dx \\
=&\int_{\mathcal{X}} \left| \frac{\log f(x) - \log g(x)}{f(x) - g(x)} \right| |f(x) - g(x)| dx \\
\leq&\frac{1}{\gamma} \int_{\mathcal{X}} |f(x) - g(x)|\, dx \\
\leq&\frac{1}{\gamma}\xi.
\end{aligned}$$

□

Claim B.5. *Let $f(x), g(x) \in [\gamma, 1]$ where $\gamma > 0$. If there exists $\xi > 0$ such that $\int_{\mathcal{X}} |\log f(x) - \log g(x)|\, dx \leq \xi$, then there holds*

$$\int_{\mathcal{X}} |f(x) - g(x)|\, dx \leq \xi.$$

Proof. Using the property of the function $\phi(x) = \log x$ (as shown in Figure 4): if $x \in (0, 1]$, then the slope of a line with any two points on the function $\phi(x)$ is bounded.

In particular, we have

$$\begin{aligned}
&\int_{\mathcal{X}} |f(x) - g(x)|\, dx \\
=&\int_{\mathcal{X}} \left| \frac{\log f(x) - \log g(x)}{f(x) - g(x)} \right|^{-1} |\log f(x) - \log g(x)| dx \\
\leq&\int_{\mathcal{X}} |\log f(x) - \log g(x)|\, dx \\
\leq&\xi.
\end{aligned}$$

□

B.7 Extension of Theorem 5.2

This additional theoretical result follows Corollary B.1 in Appendix B.5.

Corollary B.2. *Given $F^\star$, F_w and F_{sw} defined above. Let $d_{\mathcal{P}}$ be the output distribution divergence and consider W2SG in Equation (7) using forward KL divergence loss. Consider $\mathcal{F}_s$ that satisfies Assumption 5.1. If the weighted Itakura–Saito divergence $\mathrm{D}_{\mathrm{WIS}}(F^\star \| F_{sw}) \leq 0$ with the weight function $w = F_{sw} - F_w$, then we have that with probability at least $1-\delta$ over the draw of n i.i.d. samples,*

$$d_{\mathcal{P}}(\hat{F}_{sw}, F^\star) \leq d_{\mathcal{P}}(F_w, F^\star) - d_{\mathcal{P}}(F_w, \hat{F}_{sw}) + \mathcal{O}(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right), \quad (49)$$

where $\mathcal{C}_{\mathcal{F}_s}$ is a constant capturing the complexity of the function class $\mathcal{F}_s$, and the asymptotic notation is with respect to $\varepsilon \to 0, n \to \infty$.

Proof. For a clear presentation, let

$$\begin{aligned}
A &= d_{\mathcal{P}}(F_{sw}, F_s) \\
B &= d_{\mathcal{P}}(F_w, F_{sw}) \\
C &= d_{\mathcal{P}}(F_w, F_s) \\
D &= d_{\mathcal{P}}(F_s, F^\star) = \varepsilon \\
E &= d_{\mathcal{P}}(F_{sw}, F^\star) \\
F &= d_{\mathcal{P}}(F_w, F^\star) \\
G &= d_{\mathcal{P}}(\hat{F}_{sw}, F^\star) \\
H &= d_{\mathcal{P}}(F_{sw}, \hat{F}_{sw}) \\
I &= d_{\mathcal{P}}(F_w, \hat{F}_{sw}).
\end{aligned}$$

The main proof idea follows Appendix B.6. Specifically, we first replace $F^\star$ with F_s in the final step of proof of Corollary B.1, we obtain

$$C \geq A + B. \quad (50)$$

Notice that

$$E = A + D - \underbrace{\int_{\mathcal{X}} (F_{sw}(x) - F_s(x)) \log \frac{F^\star(x)}{F_s(x)} dx}_{t_1}, \quad (51)$$

$$F = C + D - \underbrace{\int_{\mathcal{X}} (F_w(x) - F_s(x)) \log \frac{F^\star(x)}{F_s(x)} dx}_{t_2}, \quad (52)$$

$$G = E - H - \underbrace{\int_{\mathcal{X}} (\hat{F}_{sw}(x) - F_{sw}(x)) \log \frac{F^\star(x)}{F_{sw}(x)} dx}_{t_3}. \quad (53)$$

Combining (50) and (51), we get

$$E \leq C + D - B - t_1. \quad (54)$$

According to Lemma B.4, we have that with probability at least $1-\delta$ over the draw of $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ that were used to construct $\hat{F}_{sw}$,

$$I \leq B + \underbrace{\mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right)}_{t_4} + \underbrace{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)}_{t_5}. \quad (55)$$

Combining (54) with (55) and we have

$$E \leq C + D - I - t_1 + t_4 + t_5. \tag{56}$$

Combining (52) with (56) and we have

$$E \leq F - I - t_1 + t_2 + t_4 + t_5. \tag{57}$$

Combining (53) with (57) and we have

$$G \leq F - I - H - t_1 + t_2 - t_3 + t_4 + t_5. \tag{58}$$

We replace $F^\star$ with $\hat{F}_{sw}$ in the final step of proof of Corollary B.1 (Recall the fact that $\hat{F}_{sw} \in V_s$ and (27): substituting $\hat{F}_{sw}$ for g in (19), and using (27)), we obtain:

$$I \geq H + B. \tag{59}$$

Combining (59) with (55) and we have

$$0 \leq H \leq t_4 + t_5 = \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{60}$$

Combining (60) with (58) and we have

$$G \leq F - I - t_1 + t_2 - t_3 + t_4 + t_5. \tag{61}$$

While t_4 and t_5 are known in (55), we analyze t_1, t_2 and t_3 one by one.

Deal with t_1. We know that

$$t_1 = \int_{\mathcal{X}} (F_{sw}(x) - F_s(x)) \log \frac{F^\star(x)}{F_s(x)} dx.$$

Using the fact that $|F_{sw}(x) - F_s(x)| \leq 1$, we have

$$|t_1| \leq \int_{\mathcal{X}} \left| \log \frac{F^\star(x)}{F_s(x)} \right| dx = \int_{\mathcal{X}} |\log F_s(x) - \log F^\star(x)| \, dx. \tag{62}$$

According to Pinsker's inequality,

$$\int_{\mathcal{X}} |F_s(x) - F^\star(x)| \, dx \leq \sqrt{\frac{1}{2} \mathrm{d}_{\mathrm{KL}}(F_s \| F^\star)} = \sqrt{\frac{1}{2}\varepsilon}. \tag{63}$$

Substitute $f(x) = F_s(x)$, $g(x) = F^\star(x)$ and $\xi = \sqrt{\frac{1}{2}\varepsilon}$ into Claim B.4 and recall (62), we have

$$|t_1| \leq \frac{1}{\gamma}\sqrt{\frac{1}{2}\varepsilon} = O(\sqrt{\varepsilon}). \tag{64}$$

Deal with t_2. The proof for t_2 is similar for t_1. In particular, replacing F_{sw} with F_w in the above and we can get

$$|t_2| = O(\sqrt{\varepsilon}). \tag{65}$$

Deal with t_3. We know that

$$t_3 = \int_{\mathcal{X}} (\hat{F}_{sw}(x) - F_{sw}(x)) \log \frac{F^\star(x)}{F_{sw}(x)} dx.$$

According to Lemma B.4, with probability at least $1-\delta$ over the draw of $(x_1, y_1), \ldots, (x_n, y_n)$, we have

$$\left| d_{\mathcal{P}}(F_w, \hat{F}_{sw}) - d_{\mathcal{P}}(F_w, F_{sw}) \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{66}$$

Notice that

$$\begin{aligned} H &= d_{\mathcal{P}}(F_{sw}, \hat{F}_{sw}) \\ &= d_{\mathcal{P}}(F_w, F_{sw}) - d_{\mathcal{P}}(F_w, \hat{F}_{sw}) + \int_{\mathcal{X}} (F_w(x) + F_{sw}(x)) \log \frac{F_{sw}(x)}{\hat{F}_{sw}(x)} dx. \end{aligned} \tag{67}$$

Substitute (60) and (66) into Equation (67) with the triangle inequality for absolute values, we get

$$\left| \int_{\mathcal{X}} (F_w(x) + F_{sw}(x)) \log \frac{F_{sw}(x)}{\hat{F}_{sw}(x)} dx \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Since $|F_w(x) + F_{sw}(x)|$ is bounded, we have

$$\left| \int_{\mathcal{X}} \left[\log F_{sw}(x) - \log \hat{F}_{sw}(x) \right] dx \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Using Claim B.5, we have

$$\left| \int_{\mathcal{X}} (\hat{F}_{sw}(x) - F_{sw}(x)) dx \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Since $\left| \log \frac{F^\star(x)}{F_{sw}(x)} \right|$ is bounded, there holds

$$|t_3| = \left| \int_{\mathcal{X}} (\hat{F}_{sw}(x) - F_{sw}(x)) \log \frac{F^\star(x)}{F_{sw}(x)} dx \right| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{68}$$

Therefore, combing (64), (65) and (68), we have

$$|t_1| + |t_2| + |t_3| \leq O(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{69}$$

Finally, combing (55) and (61) with (60) and (69), we get the result:

$$d_{\mathcal{P}}(\hat{F}_{sw}, F^\star) \leq d_{\mathcal{P}}(F_w, F^\star) - d_{\mathcal{P}}(F_w, \hat{F}_{sw}) + O(\sqrt{\varepsilon}) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}}}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where in the last inequality, we instantiate asymptotics with respect to $\varepsilon \to 0$ and $n \to \infty$. □

C Further Details and Results of Experiments

C.1 Training Details of Experiments in Language Models

The dataset is randomly divided into three distinct subsets:

- 4K samples (ground truth): They are used to fine-tune weak and strong base language models;
- 4K samples (held-out set): These labels are predicted by the weak model and used to provide weak supervision for training the strong model;
- 4K samples (the remaining): They are used for testing and evaluating the performance of all models.

When fine-tuning the weak-to-strong models, we follow (Yang et al., 2025) to set the batch size to 32, learning rate to 10^{-5}, `max_seq_len` to 512. The training epoch is set to 1 to avoid overfitting. All experiments are conducted on NVIDIA A100 80G.

C.2 Weak-to-Strong Training Procedure in Synthetic Experiments

We explore two methods to generate the weak and strong representations, which follows the experimental setting in (Charikar et al., 2024).

- **Pre-training.** We begin by randomly sampling T fine-tuning tasks $f_1^\star, \ldots, f_T^\star \in \mathcal{F}_s$. For each $t \in \{1, \cdots, T\}$, we generate N_r data $\{x_j^{(t)}\}_{j=1}^{N_r}$ where $x_j^{(t)} \sim \mathcal{P}$. Let the representations $h_w, h_s : \mathbb{R}^8 \to \mathbb{R}^{16}$ be 2-layer and 8-layer MLP with ReLU activations, respectively. And $h_w \in \mathcal{H}_w$, $h_s \in \mathcal{H}_s$. We obtain h_w and h_s via gradient descent on the representation parameters to find the minimizer of output distribution divergence loss. Specifically, We use Equation (4) as the loss function on T tasks:

$$h_l = \text{argmin}_{h \in \mathcal{H}_l} \frac{1}{T} \sum_{t=1}^{T} \hat{d}_{\mathcal{P}}(f_t^\star \circ h, f_t^\star \circ h^\star), \tag{70}$$

  where $l \in \{w, s\}$, $T = 10$, and $N_r = 2000$. Additionally, the realizable setting (Corollary B.1) is considered by explicitly setting $h_s = h^\star$, and only obtaining h_w as above.

- **Perturbations.** As an alternative, we directly perturb the parameters of $h^\star$ to obtain the weak and strong representations. Specifically, we add independent Gaussian noise $\mathcal{N}(0, \sigma_s^2)$ to every parameter in $h^\star$ to generate h_s. Similarly, we perturb $h^\star$ with $\mathcal{N}(0, \sigma_w^2)$ to generate h_w. To ensure the strong representation h_s is closer to $h^\star$ than h_w, we set $\sigma_s = 0.1$ and $\sigma_w = 9$.

Weak Model Fine-tuning. After obtaining h_w and h_s, we fix these representations and train weak models on new fine-tuning tasks. We randomly sample M new fine-tuning tasks $f_1^\star, \ldots, f_M^\star \in \mathcal{F}_s$, and generate data $\{x_j^{(i)}\}_{j=1}^{N_f}$, where $x_j^{(i)} \sim \mathcal{P}$. For each task $i = \{1, \cdots, M\}$, the weak model is trained through:

$$f_w^{(i)} = \text{argmin}_f \frac{1}{M} \sum_{i=1}^{M} \hat{d}_{\mathcal{P}}(f_t^\star \circ h^\star, f \circ h_w), \tag{71}$$

where $M = 100, N_f = 2000$. Here, the representation parameters h_w are frozen, and $f_w^{(i)}$ is learned via gradient descent. Weak models are thus trained on true data.

Weak-to-Strong Supervision. Using the trained weak models, we generate weakly labeled datasets for each fine-tuning task. Specifically, for each $i \in \{1, \cdots, M\}$, we generate $\{\tilde{x}_j^{(i)}\}_{j=1}^{N_f}$ where $\tilde{x}_j^{(i)} \sim \mathcal{P}$. The strong models are then trained on these weakly labeled datasets by solving the following optimization problem using reverse KL divergence loss for each task $i \in \{1, \cdots, M\}$:

$$f_{sw}^{(i)} = \text{argmin}_{f \in \mathcal{F}} \hat{d}_{\mathcal{P}}(f \circ h_s, f_w^{(i)} \circ h_w). \tag{72}$$

At this stage, the weak-to-strong training procedure is complete.