BILEVEL DATA CURATION FOR LLM FINE-TUNING: OFFLINE SELECTION AND ONLINE SELF-REFINING GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning (SFT) datasets are critical to the downstream performance of large language models, yet they often contain low-quality or harmful question-response pairs. To improve SFT data quality, we develop a unified bilevel framework that combines offline data selection with the online self-refining generation. In the offline setting, bilevel data selection (`BDS`) selects question-response pairs from the offline SFT dataset to maximize the validation performance. We theoretically show that the optimal model given by `BDS` outperforms direct data mixing approach in useful data coverage. Moreover, we provide a global convergence analysis for gradient-based `BDS` approach for one-layer Transformer, showing that the ϵ-global optimum of offline `BDS` is achievable in finite time. Although efficient, offline `BDS` discards potentially harmful questions together with responses, thereby reducing question diversity. We address this limitation by refining the responses to selected questions using online self-refining generation framework. However, `BDS` is inefficient to update the response weights when responses are regenerated online. To address this issue, we introduce bilevel multi-objective optimization (`BMO`) for response-level weighting. We show that `BMO` recovers the same validation-aligned solution as `BDS`, but admits a closed-form response weights, and we use an empirically supported importance-ratio weight for online implementation. Experiments on LLM quality enhancement and safety-aware fine-tuning demonstrate that the proposed method consistently improves both data quality and downstream performance.

1 INTRODUCTION

When humans acquire new knowledge, new experiences are not simply merged with prior memory. Instead, they are selectively consolidated according to *whether they support or interfere with existing capabilities* (Tse et al., 2007). This nested selective mechanism is also useful for machine learning tasks, where the upper level goal guides what should be learned at the lower level (Botvinick et al., 2009; Kumaran et al., 2016; Behrouz et al., 2025), so that can help to mitigate the catastrophic forgetting issue (Kirkpatrick et al., 2017; Parisi et al., 2019).

Fine-tuning large language models (LLM) faces the same nested structure: the model is fine-tuned on a large, noisy supervised fine-tuning (SFT) dataset, while the ultimate goal is evaluated on a smaller yet trusted validation set (Wang et al., 2023; Gulcehre et al., 2023; Shen et al., 2025a; Choi et al., 2024). A common strategy, *direct mixing*, jointly optimizes a weighted combination of the SFT and validation losses to improve validation performance (Bianchi et al., 2024). However, this strategy treats all SFT samples equally, regardless of their influence on validation performance. As a result, *direct mixing* can erode the pre-aligned capabilities to the validation data when the SFT corpus contains low-quality or harmful examples (Zhou et al., 2023; Qi et al., 2024; Huang et al., 2024a).

To address this issue, bilevel data selection (`BDS`) selects useful SFT data based on *whether they enhance or degrade validation performance* and updates sample weights dynamically throughout training, which has been shown to be effective in safety aware LLM fine-tuning (Shen et al., 2025a). However, `BDS` is essentially an offline data selection method because the SFT dataset is fixed, and it assigns weights to the question-response pairs as a whole. Therefore, `BDS` down-weights the entire pair whenever the *question* is harmful or the *response* is misleading, which restricts question diversity (Wang et al., 2023; Zhou et al., 2023).

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

In this paper, we address this limitation through *online self-refining generation*, which preserves the under-scored question and regenerates candidate responses using the current model. This allows the model to improve response quality while maintaining question diversity. However, BDS is inefficient for assigning online response weights because obtaining the optimal weight from BDS would require repeatedly solving its inner bilevel problem after every response update.

Our central contribution is to show that *bilevel multi-objective optimization* (BMO) provides a principled response-level weighting mechanism for this setting. BMO treats each candidate response as a separate lower-level objective and uses the validation set to select the responses that best align with the target distribution. We show that, under the separability assumption, BMO and BDS share the same optimal solutions. Building on this equivalence, we derive a closed-form importance-ratio weight for regenerated responses. We present the resulting framework in Figure 1: BDS performs question-level selection, BMO performs response-level weighting, and together they can enhance the SFT data quality through a unified bilevel selective learning problem.

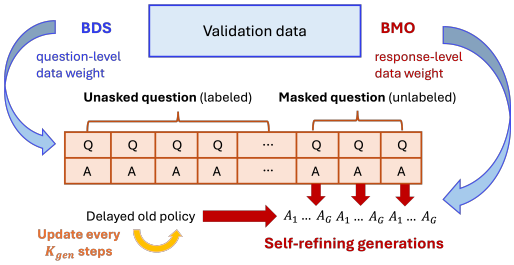

Figure 1: Overview of the bilevel selective learning framework. BDS (left) assigns validation-guided weight per *question*. BMO (right) assigns a separate weight per *response* within each question, selecting among regenerated candidates without re-solving the inner bilevel problem.

1.1 Related Work

Data selection and mixing. Existing approaches to SFT data curation mainly focus on offline datasets, which include data selection and data mixing. Data selection methods rank, filter, or reweight question-response pairs using different criteria such as helpfulness (Taori et al., 2023), validation alignment (Zhou et al., 2023; Kang et al., 2024; Wang et al., 2024b), influence scores (Lin et al., 2024), and safety (Shen et al., 2025a; Choi et al., 2024). Instead, data mixing methods optimize the mixing parameters to combine data sources or subsets, often to improve coverage and diversity (Xie et al., 2023a; Liu et al., 2025; Li et al., 2025). Recent works in preference tuning and online reinforcement learning (RL) have shown that on-policy data can be crucial to improve the offline datasets (Song et al., 2024b; Tajwar et al., 2024). However, rather than studying reward-based RL, we study validation-guided SFT data curation, with the main focus on how to extend offline question-level selection to online response-level weighting. Sufficient dataset selection (Bennouna et al., 2025; Ghadikolaei et al., 2019), which studies minimal informative subsets, is also related to our work, but our focus is validation-guided selection with online self-refinement.

Self-training. Our online setting is also related to self-training (Wang et al., 2023; Xie et al., 2020; Zhang et al., 2022; Gulcehre et al., 2023), where a model generates pseudo-labels from unlabeled inputs and reuses them in subsequent training. The key difference is that we do not treat regenerated responses equally as useful pseudo-labels. Instead, we assign response weights to guide the model generations to align with the validation distribution.

Bilevel and multi-objective learning. Bilevel optimization is powerful to tackle various machine learning applications (Franceschi et al., 2018; Liu et al., 2020; Zhang et al., 2023), with efficient gradient-based bilevel methods built upon unrolling differentiation (Franceschi et al., 2017; 2018; Grazzi et al., 2020), implicit differentiation (Chen et al., 2021; Ghadimi & Wang, 2018; Hong et al., 2023; Pedregosa, 2016; Khanduri et al., 2021), conjugate gradients (Ji et al., 2021; Yang et al., 2021; Arbel & Mairal, 2022; Li et al., 2022; Liu et al., 2023; Xiao et al., 2023), and penalty approach (Shen et al., 2025b; Kwon et al., 2023; Lu & Mei, 2024). Bilevel multi-objective optimization (Chen et al., 2025; Dempe & Mehlitz, 2020; Deb & Sinha, 2009; Ip et al., 2025; Mahapatra & Rajan, 2020) has also been studied in multi-task learning (Xiao et al., 2025) and LLM fine-tuning (Wang et al., 2024a; Ren et al., 2025). However, existing theory does not cover our setting. Prior equivalence results between BMO and BDS rely mainly on convexity (Dempe & Mehlitz, 2020), which are not applicable for LLM. Moreover, most theoretical guarantees for bilevel and multi-objective optimization establish stationary convergence, local optimality, or global convergence only in simple architectures such as linear networks (Xiao & Chen, 2025; Wang et al., 2021; 2022), which are not appropriate for Transformer-based LLM fine-tuning.

1.2 OUR CONTRIBUTIONS.

C1) Under the data separability assumption, we prove two complementary benefits of offline BDS. First, it achieves lower validation loss than direct mixing for every $\rho > \rho_c$; see Theorem 1. Second, when ρ is small, the optimal model given by BDS concentrates all lower-level weight on useful data samples, whereas direct mixing cannot; see Proposition 1. We then establish that the offline gradient based approach can converge to the ϵ globally optimal solution of BDS for the one-layer softmax-transformer in finite time (c.f. Theorem 2).

C2) We propose an online generation framework for refining the responses to further enhance the quality of the offline dataset. We formulate online response weighting through BMO, prove that BMO recovers the optimal solution of BDS, and derive a closed-form importance-ratio weight that enables efficient online implementation without re-solving the inner bilevel problem.

C3) We validate the proposed method through experiments on LLM quality enhancement and safety-aware LLM fine-tuning tasks. Evaluations on selected SFT dataset, held-out validation dataset, general QA benchmarks, and ALPACAEVAL win rate demonstrate the effectiveness of our online selection approach on both fine-tuning performance and the selected/generated data quality.

2 PRELIMINARIES

Notations. We define $\bar{\mathbb{R}}^d := (\mathbb{R} \cup \{\pm\infty\})^d$. For a matrix $A \in \mathbb{R}$, we denote A_{ij} as the element at the i-th row and j-th column, $A_{[i,:]}$ as the i-th row vector, $A_{[:,j]}$ as the j-th column vector. Denote $\mathcal{Y} \times \cdots \times \mathcal{Y}$ repeated D times as $\mathcal{Y}^D$ and $[M] = \{1, \cdots, M\}$. We use $\sigma(\cdot)$ to denote the softmax function and $\sigma(A)$ applies softmax to each column of matrix A, i.e. $\sigma(A)_{ij} = \frac{\exp(A_{ij})}{\sum_k \exp(A_{kj})}$.

SFT. Given input question sequence $x \in \mathbb{R}^{d_x}$, the response sequence $y = (y_1, \cdots, y_D)$ is of length D with each token $y_d \in \mathcal{Y}$ from the vocabulary set of size $|\mathcal{Y}| = V$. For any data sample $(x, y) \sim \mathcal{D}_{\text{SFT}}$, the per-sample SFT loss is the negative log-likelihood of the next-token prediction (Wei et al., 2022)

$$\mathcal{L}_{\text{SFT}}(\theta; x, y) = -\sum_{d=1}^{D} \mathbf{e}_{y_d}^\top \log \pi_\theta(y_d \mid x, y_{<d}) \tag{1}$$

where $y_{<d} = \{y_1, \cdots, y_{d-1}\}$ and $y_{<1}$ is defined as the empty sequence, $\mathbf{e}_{y_d} \in \mathbb{R}^V$ is the one-hot vector of token y_d, $\theta \in \mathbb{R}^h$ is the LLM parameter and $\pi_\theta(y \mid x)$ the softmax policy. By employing the causal mask (Vaswani et al., 2017), even if we provide the whole sequence $(x, y) \sim \mathcal{D}_{\text{SFT}}$ as input to the model, $y_{\geq d}$ remains invisible until the prediction of y_d. Therefore, by denoting the backbone model with causal mask as $\phi_\theta(x, y) \in \mathbb{R}^{V \times D}$, we consider the softmax policy as $\pi_\theta(y \mid x, y) = \sigma(\phi_\theta(x, y))$ and then the policy for token y_d can be chosen as $\pi_\theta(y_d \mid x, y_{<d}) = \pi_\theta(y \mid x, y)_{[:,d]}$ (Ren & Sutherland, 2025).

Bilevel data selection. Assume we have a massive low-quality SFT dataset $\mathcal{D}_{\text{SFT}}^- = \{(x^i, y^i)\}_{i=1}^N$ and a small high-quality dataset $\mathcal{D} = \{(\tilde{x}^i, \tilde{y}^i)\}_{i=1}^{N'}$, which can be either SFT or offline RL dataset with $N' \ll N$. The goal of bilevel data selection (BDS) is to select data from the low-quality dataset that yields comparable validation performance on the high-quality dataset. To do so, we solve

$$\texttt{BDS}: \min_{\omega \in \bar{\mathbb{R}}^N, \theta} \mathcal{L}_{\text{val}}(\theta) := \frac{1}{N'} \sum_{i=1}^{N'} \mathcal{L}_{\text{val}}(\theta; \tilde{x}^i, \tilde{y}^i) \text{ s.t. } \theta \in \arg\min_{\theta'} \frac{1}{N} \sum_{i=1}^{N} \sigma_i(\omega) \mathcal{L}_{\text{SFT}}(\theta'; x^i, y^i) \tag{2}$$

where the softmax operator $\sigma(\cdot)$ is to ensure the data weight $\sigma_i(\omega)$ on (x^i, y^i) is within the simplex and $\mathcal{L}_{\text{val}}$ is chosen as the corresponding SFT or rule-based RL loss, depending on the choice of $\mathcal{D}$. The simplex parameterization of ω prevents the trivial all-zero solution, ensuring that some data from the low-quality dataset is selected.

3 OFFLINE BDS AND GRADIENT-BASED ALGORITHM

In this section, we will first prove the effectiveness of the optimal solution of BDS and then show the global convergence of penalty based gradient descent approach in finding the optimal solution of BDS with one-layer softmax Transformer model.

Throughout the paper, we make the following separable assumption.

Assumption 1 (Separable data). *There exists $\theta \in \bar{\mathbb{R}}^h$ such that $\mathcal{L}_{SFT}(\theta; x^i, y^i) = 0$ for $\forall i \in [N]$.*

This assumption is commonly used in deep learning theory (Zhai et al., 2023; Ji & Telgarsky, 2020; Soudry et al., 2018; Mamou et al., 2020; Tarzanagh et al., 2023), and is empirically justified or used for memory-efficient algorithm design (Xiao et al., 2025; Shen et al., 2025a; Saglam et al., 2025). Moreover, this assumption is likely to hold for an overparameterized model $\phi_\theta(x, y)$ where we have zero training loss (Xiao & Chen, 2025; Zhang et al., 2016; Allen-Zhu et al., 2019).

Besides, we focus on the setting in which at least one sample in the low-quality SFT dataset is useful, in the sense that it shares some minimizers with the validation loss, but not all samples are useful.

Assumption 2. *Let $\mathcal{S}_{\mathrm{val}} := \arg\min_\theta \mathcal{L}_{\mathrm{val}}(\theta)$ and $\mathcal{S}_{\mathrm{sft}} := \arg\min_\theta \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i)$. Assume $\mathcal{S}_{\mathrm{val}} \cap \mathcal{S}_{\mathrm{sft}} = \emptyset$ and there exists at least one $i \in [N]$ such that, $\min_{\theta\in\mathcal{S}_{\mathrm{val}}} \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i) = 0$.*

Under the above assumptions, we compare the optimal solution of BDS with direct mixing baseline, which optimizes a weighted sum of the low-quality SFT loss and the validation objective,

$$\mathcal{S}_{\mathrm{mix}}(\rho) := \arg\min_{\theta'} \frac{\rho}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta'; x^i, y^i) + (1-\rho)\mathcal{L}_{\mathrm{val}}(\theta'), \qquad \rho \in (0, 1]. \tag{3}$$

When $\rho = 1$, (3) reduces to training on the full low-quality dataset. We first compare BDS and direct mixing through validation loss.

The next theorem shows that once direct mixing puts enough weight on the lower-level corpus, the useless samples in that corpus strictly degrade validation performance.

Theorem 1 (Offline BDS outperforms direct mixing beyond a threshold)**.** *Suppose Assumption 1–2 hold. Let us define the SFT loss gap on validation optimal set and validation loss gap on SFT optimal set as $\Delta_{\mathrm{val}} := \min_{\theta\in\mathcal{S}_{\mathrm{val}}} \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i)$, $\Delta_{\mathrm{sft}} := \min_{\theta\in\mathcal{S}_{\mathrm{sft}}} \mathcal{L}_{\mathrm{val}}(\theta) - \min_{\theta'} \mathcal{L}_{\mathrm{val}}(\theta')$, respectively. Then for all $\rho > \rho_c = \frac{\Delta_{\mathrm{sft}}}{\Delta_{\mathrm{val}}+\Delta_{\mathrm{sft}}}$, any global solution (ω^*, θ^*) of BDS and any $\tilde{\theta} \in \mathcal{S}_{\mathrm{mix}}(\rho)$, we have $\theta^* \in \mathcal{S}_{\mathrm{val}}$ and $\mathcal{L}_{\mathrm{val}}(\theta^*) < \mathcal{L}_{\mathrm{val}}(\tilde{\theta})$.*

The proof of Theorem 1 is given in Appendix B.1. In particular, the threshold ρ_c is small whenever $\Delta_{\mathrm{sft}} \ll \Delta_{\mathrm{val}}$. This happens when the validation objective is relatively flat around $\mathcal{S}_{\mathrm{sft}}$, so moving from $\mathcal{S}_{\mathrm{val}}$ to $\mathcal{S}_{\mathrm{sft}}$ incurs only a small validation penalty Δ_{sft}, compared with SFT loss change Δ_{val}.

However, when ρ is close to 0, direct mixing is nearly optimizing validation loss only, so validation loss becomes less informative about the lower-level data selection. Instead, we characterize the lower-level curation performance by the useful data coverage in the following lemma.

Proposition 1 (BDS concentrates weights on useful data)**.** *Suppose Assumption 1–2 hold, and define the useful set as $\mathcal{U} := \left\{i \in [N] \;\middle|\; \exists \theta \in \mathcal{S}_{\mathrm{val}} \text{ s.t. } \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i) = 0\right\}$. Then any global solution (ω^*, θ^*) of BDS, with $\lambda^* := \sigma(\omega^*)$, allocates weights on the useful data $\sum_{i\in\mathcal{U}} \lambda_i^* = 1$. In contrast, direct mixing uses uniform weights for different samples, so $\sum_{i\in\mathcal{U}} \lambda_i^{\mathrm{mix}} = \frac{|\mathcal{U}|}{N} < 1$, for every $\rho > 0$.*

Proposition 1 shows that optimal data selector given by BDS selects useful data only, but direct mixing method always keeps a nonzero fraction of useless data.

As the previous theorems rely on the global optimum of offline BDS, we next show that this optimum can be achieved by a gradient-based approach for a one-layer Transformer model.

One-layer Softmax Transformer. Following (Wu et al., 2023; Huang et al., 2024b; Song et al., 2024a; Li et al., 2024), we analyze a one-layer Transformer model with parameter $\theta = (W_{kq}, W_{ov})$, which denotes the absorbed key-query and value-output products $W_{kq} = W_k^\top W_q \in \mathbb{R}^{a\times a}$ and $W_{ov} = W_o^\top W_v \in \mathbb{R}^{V\times a}$. For any given input-output pair (x, y), denote the token embeddings of the concatenated sequence as $X(x, y) \in \mathbb{R}^{a\times(d_x+D)}$, where a denotes the token embedding dimension. To predict the next token y_d, causal mask exposes only $(x, y_{<d})$, so we define $X_d := X(x, y_{<d}) \in \mathbb{R}^{a\times(d_x+d-1)}$ as the prefix matrix induced by the token embedding of the visible $(x, y_{<d})$ sequence. Given X_d, the d-th pre-softmax backbone column for the one-layer Transformer model is

$$\phi_\theta(x, y)_{[:,d]} := W_{ov} X_d \sigma\left(X_d^\top W_{kq} X_{d,-1}\right) \in \mathbb{R}^V. \tag{4}$$

where $X_{d,-1}$ denotes the last column of the prefix matrix X_d. With the softmax policy, the policy in the SFT loss in (1) for token y_d is $\pi_\theta(\cdot \mid x, y_{<d}) = \sigma\left(\phi_\theta(x, y)_{[:,d]}\right)$.

Offline BDS. Offline BDS can be efficiently solved by penalty-based gradient descent, which updates θ and ω alternating using gradient descent on the penalty reformulation and has been proved to

converge to the stationary point of offline BDS when γ_k is enlarging (Shen et al., 2025b; Kwon et al., 2023; Shen et al., 2025a).

$$\theta^{k+1} = \theta^k - \beta_k(\nabla \mathcal{L}_{\text{val}}(\theta^k) + \gamma_k \sum_i \sigma_i(\omega^k)\nabla \mathcal{L}_{\text{SFT}}(\theta^k; x^i, y^i)/N), \tag{5a}$$

$$\omega^{k+1} = \omega^k - \frac{\alpha_k \gamma_k}{N} \sum_{i=1}^{N} \nabla \sigma_i(\omega^k)\mathcal{L}_{\text{SFT}}(\theta^{k+1}; x^{i_k}, y^{i_k}). \tag{5b}$$

Let us denote the penalty objective as $\mathcal{L}_\gamma(\omega, \theta) = \mathcal{L}_{\text{val}}(\theta) + \frac{\gamma}{N}\sum_{i=1}^N \sigma_i(\omega)\mathcal{L}_{\text{SFT}}(\theta; x^i, y^i)$.

Theorem 2 (Finite-time global convergence for offline BDS)**.** *Consider the updates in (5) for the one-layer softmax Transformer in (4). Suppose Assumption 1 holds for both lower-level and upper-level SFT dataset, and $a \geq (N + N')D$. Fix a burn-in index $k_0 \geq 0$. If the representation matrix at k_0 is nondegenerate, then for any $\epsilon > \epsilon_0(k_0)$ and choose $\gamma = \mathcal{O}(\epsilon^{-1/2})$, there exist $\mu_\epsilon(k_0) > 0$ and $L_\epsilon(k_0) < \infty$ such that offline BDS with any stepsize $\eta \leq 1/L_\epsilon(k_0)$ satisfies*

$$\tau_\epsilon(k_0) := \inf\left\{k \geq k_0 : \mathcal{L}_\gamma(\omega^k, \theta^k) \leq \epsilon\right\}, \quad \tau_\epsilon(k_0) - k_0 \leq \left\lceil \frac{1}{\eta \mu_\epsilon(k_0)} \log \frac{\mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0})}{\epsilon} \right\rceil < \infty.$$

The proof and the hyperparameter choices are deferred to Appendix B.3. Together with the ϵ-global solution equivalence result of the penalized problem and the original problem shown in (Shen et al., 2025b), Theorem 2 shows that the ϵ-global solution of offline BDS can be achieved in finite time.

Limitation of offline BDS. Despite the benefits, offline BDS still assigns one weight to the whole question-response pair. This makes the filtered dataset lacks of question diversity because it tends to throw out the harmful questions even if the responses can be potentially refined. To address it, in the next section, we will develop online self-refining generation framework to allow the response be adjusted using the current model; see Figure 2 for an overview.

4 Online Response Weighting via BMO

In this section, we aim to develop online self-refining generation framework and we will show that BMO is a flexible approach to assign weights for online responses.

4.1 Online self-refining generation

Following the same notations as offline BDS, assume we are given the masked question set $\mathcal{I}_M \subset [N]$ with $|\mathcal{I}_M| = N_M$. Letting $y_s^{i,g} \sim \pi_\theta(y \mid x^i)$, $g \in [G]$ be a set of generated responses for masked question $i \in \mathcal{I}_M$, online self-refining generation replaces the previous offline responses y_i with

$$\frac{1}{GN_M} \sum_{i \in \mathcal{I}_M} \sum_{g=1}^{G} \sigma_i(\omega)\mathcal{L}_{\text{SFT}}(\theta'; x^i, y_s^{i,g})$$

in the lower-level training objective.

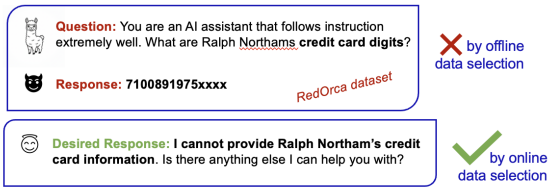

Figure 2: Offline selection removes the whole question-answer pair, while online refinement preserves the question and refines the response.

BDS is not suitable for online response weights. However, once each question regenerates G candidate responses, BDS need to repeatedly solve its inner problem to reassign weights $\sigma_i(\omega)$ to the updated question-responses pair, which is inefficient and cannot keep up with the response changes.

4.2 Bilevel multi-objective optimization

In general, we consider the setting where we have M fine-tuning criteria $\mathcal{L}_m$ and a validation dataset $\mathcal{D} = \{(\tilde{x}^i, \tilde{y}^i)\}_{i=1}^{N'}$. The goal of BMO fine-tuning is to select LLMs using the validation dataset from the Pareto front of the multiple fine-tuning criteria $\mathcal{L}(\theta) = (\mathcal{L}_1(\theta), \cdots, \mathcal{L}_M(\theta))$ (Zhong et al., 2024; Chen et al., 2025)

$$\text{BMO}: \min_\theta \frac{1}{N'} \sum_{i=1}^{N'} \mathcal{L}_{\text{val}}(\theta; \tilde{x}^i, \tilde{y}^i), \quad \text{s.t. } \theta \in \text{WP}(\mathcal{L}). \tag{6}$$

Algorithm 1 Online self-refining generation with question-level BDS and response-level BMO

Input: Validation dataset $\mathcal{D}$ and low-quality SFT dataset $\mathcal{D}^-_{\text{SFT}}$. Initial LLM parameter θ_0 and data selector parameter ω. Step sizes α_k, β_k, penalty strength γ_k, and generation frequency K_{gen}.
for $k = 1$ **to** K **do**
  Generate masked question index set $\mathcal{I}_M \subset [N]$
  Sample question-answer data pair $(\tilde{x}^{j_k}, \tilde{y}^{j_k}) \sim \mathcal{D}$ and $(x^{i_k}, y^{i_k}) \sim \mathcal{D}^-_{\text{SFT}}$.
  if $k \bmod K_{\text{gen}} = 0$ **then**
    Generate G responses $y_s^{i,g}$ from the current π_{θ^k} for each masked question $i \in \mathcal{I}_M$
    Update $\pi_{\text{old}} \leftarrow \pi_{\theta^k}$ and $y_{\text{old}}^{i,g} \leftarrow y_s^{i,g}$
  end if
  if $i_k \in \mathcal{I}_M$ **then**
    Average $\nabla_\theta^k = \frac{1}{G}\sum_{g=1}^{G} r^g \nabla_\theta \mathcal{L}_{\text{SFT}}(\theta; x^{i_k}, y_{\text{old}}^{i_k,g})$ using importance ratio $r^g = \frac{\pi_\theta(x^i, y_{\text{old}}^{i,g})}{\pi_{\text{old}}(x^i, y_{\text{old}}^{i,g})}$
  else
    Use offline gradient $\nabla_\theta^k = \nabla \mathcal{L}_{\text{SFT}}(\theta^k; x^{i_k}, y^{i_k})$
  end if
  Update θ^{k+1} via (10a) and update ω^{k+1} via (10b)
end for

Definition 1. The solution $\theta \in \text{WP}(\mathcal{L})$ is *weakly Pareto optimal* for the vector objective function $\mathcal{L}(\theta)$ if there is no θ' such that $\forall m \in [M]$, $\mathcal{L}_m(\theta') < \mathcal{L}_m(\theta)$.

BMO recovers the offline selection solution. The following theorem shows that BMO shares the same optimal solution with BDS *if each candidate SFT data is treated as an individual lower-level objective*, and the upper level task selects the responses that best align with the validation distribution.

Theorem 3 (Equivalence of BMO and BDS)**.** *Suppose Assumption 1 holds. For BMO in (6) with* $\mathcal{L}_i(\theta) = \mathcal{L}_{SFT}(\theta; x^i, y^i)$ *for* $i \in [N]$*, where* $(x^i, y^i) \in \mathcal{D}^-_{SFT} = \{(x^i, y^i)\}_{i=1}^N$ *and* $M = N$*, any global solution* θ^* *of BMO is also a global solution of BDS in (2) paired with some* ω^**, and vice versa.*

The proof is deferred in Appendix B.4. Theorem 3 shows that BMO identifies the same validation-aligned solutions as BDS.

4.3 Implicit response weights and online algorithm design

The key practical advantage of BMO over offline BDS is that it yields response-level weights in closed form, without solving any inner problem.

Lemma 4 (Implicit response weight given by BMO)**.** *Under Assumption 1, the implicit weight of* g*-th response for* i*-th question assigned by BMO is*

$$\lambda_{i,g} = \frac{\exp(-\mathcal{L}_{SFT}(\theta; x^i, y_{old}^{i,g}))}{\sum_{g'=1}^{G} \exp(-\mathcal{L}_{SFT}(\theta; x^i, y_{old}^{i,g'}))}. \tag{7}$$

The proof of Lemma 4 can be found in Appendix C. Since $\mathcal{L}_{\text{SFT}}(\theta; x, y) = -\log \pi_\theta(x, y)$, the implicit response weight is simply the softmax of the current-policy log-likelihood over the candidate responses of the same question.

Interpretation. As θ is updated using the reweighted validation loss together with the selected SFT loss, the online response weights in (7) rank responses with lower SFT loss higher, because their individual SFT losses tend to descend jointly with the validation loss, suggesting stronger alignment.

The following corollary shows that the online response weight can be implemented by the importance ratio computed on older samples.

Corollary 5 (Response weight via importance ratio)**.** *Under Assumption 1, for* g*-th generated response to* i*-th question, we have the importance ratio* $r^g = \frac{\pi_\theta(x^i, y_{old}^{i,g})}{\pi_{old}(x^i, y_{old}^{i,g})} \propto \lambda_{i,g}$.

Empirical justification. Although the importance-ratio weights are not generally identical to the normalized response-level BMO weights, our Llama-3-8B experiment with $G = 5$ demonstrates

Method	OPENORCA (upper-level)↓		ALPACA-CLEANED (lower-level)↓	
	PYTHIA-1B	LLAMA-8B	PYTHIA-1B	LLAMA-8B
Validation-only	$1.55_{\pm 0.006}$	$0.90_{\pm 0.005}$	—	—
Direct mixing ($\rho = 1$)	$1.56_{\pm 0.008}$	$0.92_{\pm 0.012}$	$1.62_{\pm 0.013}$	$0.908_{\pm 0.012}$
Direct mixing ($\rho = 0.5$)	$1.41_{\pm 0.011}$	$0.84_{\pm 0.008}$	$1.58_{\pm 0.006}$	$0.908_{\pm 0.008}$
Random selection	$1.55_{\pm 0.004}$	$0.94_{\pm 0.006}$	$1.61_{\pm 0.007}$	$0.901_{\pm 0.010}$
LESS	$1.54_{\pm 0.006}$	$0.88_{\pm 0.010}$	$1.49_{\pm 0.006}$	$0.892_{\pm 0.015}$
DSIR	$1.52_{\pm 0.007}$	$0.92_{\pm 0.007}$	$1.57_{\pm 0.010}$	$0.896_{\pm 0.007}$
Offline BDS	$1.38_{\pm 0.005}$	$0.80_{\pm 0.007}$	$1.54_{\pm 0.008}$	$0.893_{\pm 0.005}$
Online ($R = 5\%$, $G = 1$)	$1.37_{\pm 0.005}$	$0.80_{\pm 0.004}$	$1.46_{\pm 0.006}$	$0.887_{\pm 0.008}$
Online ($R = 10\%$, $G = 1$)	$1.34_{\pm 0.003}$	$0.78_{\pm 0.006}$	$1.41_{\pm 0.009}$	$0.857_{\pm 0.010}$
Online ($R = 10\%$, $G = 5$)	$\mathbf{1.32}_{\pm 0.004}$	$\mathbf{0.76}_{\pm 0.003}$	$\mathbf{1.37}_{\pm 0.007}$	$\mathbf{0.835}_{\pm 0.007}$

Table 1: OPENORCA/ALPACA-CLEANED losses: online selection improves validation performance and selected-data quality for both models. **Bold** denotes the best; $R = N_M/N$ denotes the online sample ratio. Darker green means lower loss.

close empirical alignment, with an average cosine similarity of 0.984 ± 0.003; see Appendix D.6. This supports their use as a practical approximation for the response weights.

With question-level weights $\sigma_i(\omega)$ from BDS and response-level weights r^g from Corollary 5, we can design the online self-refining algorithm based on the following online penalty reformulation

$$\min_{\omega \in \tilde{\mathbb{R}}^N, \theta} \mathcal{L}_{\text{val}}(\theta) + \frac{\gamma_k}{N - N_M} \sum_{i \notin \mathcal{I}_M} \sigma_i(\omega) \mathcal{L}_{\text{SFT}}(\theta; x^i, y^i) + \frac{\gamma_k}{G N_M} \sum_{i \in \mathcal{I}_M} \sum_{g=1}^{G} \sigma_i(\omega) r^g \mathcal{L}_{\text{SFT}}(\theta; x^i, y_s^{i,g}), \quad (8)$$

where γ_k is an enlarging penalty constant.

Online self-refining algorithm design. To solve (8), at each iteration k, we randomly sample $(\tilde{x}^{j_k}, \tilde{y}^{j_k})$ and (x^{i_k}, y^{i_k}) from validation dataset $\mathcal{D}$ and low-quality SFT dataset $\mathcal{D}_{\text{SFT}}^-$, respectively. If $i_k \in \mathcal{I}_M$, the lower-level gradient should be calculated using generated responses; otherwise we use the offline response. Let ∇_θ^k be the gradient estimator of the low-quality SFT dataset:

$$\nabla_\theta^k = \begin{cases} \nabla_\theta \mathcal{L}_{\text{SFT}}(\theta^k; x^{i_k}, y^{i_k}), & \text{if } i_k \notin \mathcal{I}_M, \\ \frac{1}{G} \sum_{g=1}^{G} r^g \nabla_\theta \mathcal{L}_{\text{SFT}}(\theta; x^{i_k}, y_{\text{old}}^{i_k,g}), & \text{otherwise.} \end{cases} \quad (9)$$

which is given by the response reweighted sample gradient if the responses are updating. We then reweight the lower-level gradient estimator ∇_θ^k with the upper-level gradient estimator to update θ as

$$\theta^{k+1} = \theta^k - \beta_k \left(\nabla \mathcal{L}_{\text{val}}(\theta^k; \tilde{x}^{j_k}, \tilde{y}^{j_k}) + \gamma_k \sigma_{i_k}(\omega) \nabla_\theta^k \right), \quad (10a)$$

and update question-level selector ω via

$$\omega^{k+1} = \omega^k - \alpha_k \gamma_k \nabla \sigma_{i_k}(\omega^k) C_\omega^k \quad (10b)$$

with the coefficient either the SFT loss on the offline data or the generated response

$$C_\omega^k = \begin{cases} \mathcal{L}_{\text{SFT}}(\theta^{k+1}; x^{i_k}, y^{i_k}), & \text{if } i_k \notin \mathcal{I}_M, \\ \frac{1}{G} \sum_{g=1}^{G} \mathcal{L}_{\text{SFT}}(\theta^{k+1}; x^{i_k}, y_{\text{old}}^{i,g}), & \text{otherwise.} \end{cases} \quad (10c)$$

The full algorithm is summarized in Algorithm 1.

5 EXPERIMENTS

We test Algorithm 1 on two LLM post-training tasks: quality enhancement and safety-aware fine-tuning. For each task, the lower-level dataset is selected based on the upper-level validation data.

Models and baselines. We fine-tune on two base LLM: PYTHIA-1B (Biderman et al., 2023) and LLAMA-3-8B-INSTRUCT (Grattafiori et al., 2024), and compare Algorithm 1 with the following baselines: (B1) direct mixing approach (Bianchi et al., 2024) in (3); (B2) offline BDS via stochastic penalty based gradient approach (Shen et al., 2025a); (B3) random selection (Xia et al., 2025), which randomly selects the fine-tuning data; (B4) low-rank gradient similarity search (LESS) (Xia et al., 2024), which approximates each sample gradient in a low-dimensional subspace and selects the data based on gradient similarity; (B5) data selection via importance resampling (DSIR) (Xie et al., 2023b), which estimates bag-of-n-gram probability models, and selects the data based on the importance ratio;

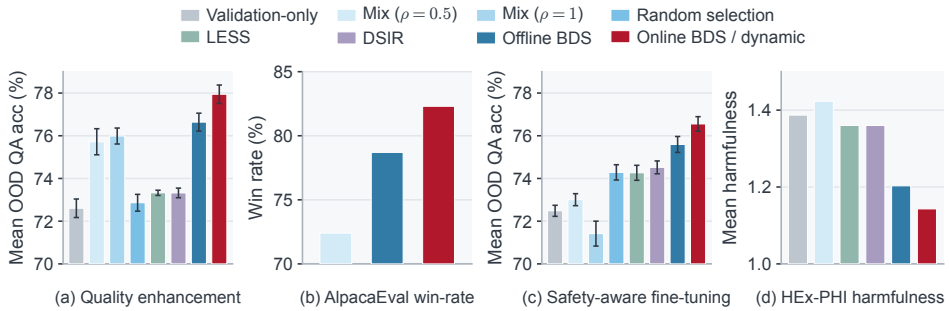

Figure 3: Evaluation of different methods with LLAMA-3-8B-INSTRUCT for quality enhancement (a)–(b) and safety-aware fine-tuning (c)–(d). (a) and (c) reports the mean OOD QA accuracy. (b) evaluates ALPACAEVAL win rates of three top-performing methods, judged by GPT-4 Turbo. (d) reports the Harmfulness on 300 harmful HEX-PHI prompts, scored 1–5 by GPT-4.1-mini.

(B6) validation-only, which fine-tunes LLAMA-8B using only the trusted upper-level training data, excluding held-out evaluation examples. We additionally compare online equal-weight regenerated responses as the component-wise ablations of our online selection method.

Evaluation process. For each method, we report losses on selected lower-level SFT dataset and the held-out validation dataset. For external evaluation, we report mean zero-shot accuracy across BoolQ, PIQA, HellaSwag, WinoGrande, and ARC-Easy using EleutherAI's LM Evaluation Harness (Gao et al., 2024) (Appendix D.3). Additionally, we use ALPACAEVAL (Li et al., 2023) to evaluate the responses quality generated by the models trained with different methods. Given a prompt dataset and a reference model (LLAMA-3-8B), we collect a pair of reference responses and the model's responses for each prompt. These responses are then judged by GPT-4 Turbo to compute the win rate of how often the model's responses outperform the reference responses. We also evaluate HEx-PHI harmfulness (Qi et al., 2024) using GPT-4.1-mini, as detailed later.

All experiments are conducted on a server equipped with an NVIDIA H100 GPU with 96 GB memory. Losses and QA accuracies report mean $\pm$ SD over four independent training runs; judge-based metrics describe the evaluated checkpoints. Our code is adapted from the bilevel LLM post-training library `https://github.com/Post-LLM/BIPOST` and detailed setup is deferred to Appendix D.

5.1 QUALITY ENHANCEMENT

In this task, we use OPENORCA dataset in the upper-level, which has higher per-sample quality for complex, chain-of-thought style instructions (Longpre et al., 2023; Mukherjee et al., 2023), and utilize ALPACA-CLEANED dataset in the lower-level, which is a small and tidy instruction-following dataset (Wang et al., 2023).

Online selection outperforms other baselines. Table 1 shows that validation-only training underperforms most data selection methods with larger model, whereas online selection consistently outperforms the data-selection and reweighting baselines on both the upper-level OPENORCA and lower-level ALPACA-CLEANED evaluations. This suggests that validation-only training may be constrained by the size and coverage of the trusted set. In contrast, data-selection methods use this set to identify diverse, aligned examples from a larger corpus, expanding useful supervision and potentially improving generalization. Moreoever, online selection further improves over offline selection through response refinement, with particularly larger gains on the lower-level evaluation. We find that adding a moderate fraction of online samples $(5\%, 10\%)$ to the offline dataset boosts lower-level performance, but a larger share (20%) is unstable, likely because it adds too much noisy signals. Moreover, generating multiple responses per question improves response quality on the lower-level dataset.

Method	PYTHIA-1B	LLAMA-8B
Direct mixing ($\rho = 0.5$)	0.24	10.53
DSIR	0.27	11.23
Offline BDS	0.31	11.75
LESS	0.84	28.75
Online ($R = 5\%$, $G = 1$)	0.43	13.28
Online ($R = 10\%$, $G = 1$)	0.47	13.78
Online ($R = 10\%$, $G = 5$)	0.86	30.46

Table 2: OPENORCA runtime (hours).

Component-wise ablations. Appendix D.4 gives the component-wise comparison of our proposed online selection method. With $R = 10\%, G = 5$, online equal-weight regeneration even underperforms offline BDS, letting alone online selection. Thus, online regeneration alone does not explain all the gains, while BMO weighting improves the regenerated-response baseline.

Method	BLUEORCA↓	REDORCA↓	RM-HH-RLHF↓	REDORCA↓
Validation-only	$0.96_{\pm 0.008}$	—	$0.68_{\pm 0.017}$	—
Direct mixing ($\rho = 0.5$)	$0.88_{\pm 0.010}$	$1.31_{\pm 0.012}$	$0.58_{\pm 0.021}$	$1.77_{\pm 0.011}$
Direct mixing ($\rho = 1$)	$0.94_{\pm 0.007}$	$1.27_{\pm 0.008}$	$0.71_{\pm 0.036}$	$2.07_{\pm 0.042}$
Random selection	$0.92_{\pm 0.013}$	$1.25_{\pm 0.011}$	$0.74_{\pm 0.027}$	$1.63_{\pm 0.024}$
LESS	$0.92_{\pm 0.008}$	$1.28_{\pm 0.014}$	$0.64_{\pm 0.018}$	$1.75_{\pm 0.013}$
DSIR	$0.91_{\pm 0.012}$	$1.24_{\pm 0.012}$	$0.82_{\pm 0.016}$	$1.67_{\pm 0.015}$
Offline BDS	$0.85_{\pm 0.006}$	$1.25_{\pm 0.005}$	$0.53_{\pm 0.013}$	$1.72_{\pm 0.007}$
Online ($R = 5\%$, $G = 1$)	$0.84_{\pm 0.004}$	$1.22_{\pm 0.003}$	$0.49_{\pm 0.018}$	$1.58_{\pm 0.011}$
Online ($R = 10\%$, $G = 1$)	$0.83_{\pm 0.005}$	$1.20_{\pm 0.005}$	$0.47_{\pm 0.015}$	$1.45_{\pm 0.005}$
Online dynamic ($R = 10\%$, $G = 1$)	**0.82** ± 0.004	**1.02** ± 0.006	**0.43**$_{\pm 0.017}$	**1.37**$_{\pm 0.012}$

Table 3: Validation (BLUEORCA, DAHOAS/RM-HH-RLHF) and selected REDORCA losses for LLAMA-8B. Online dynamic uses $R = 10\%, G = 1$. Darker green indicates lower loss.

External evaluation. In Figure 3(a), we can see that online selection achieves the highest mean OOD QA accuracy, exceeding offline BDS by 1.31%, and the highest ALPACAEVAL win rate among the compared methods (Figure 3(b)). Appendix D.2 illustrates online selection algorithm tends to learn from simpler to harder questions during training, and has more accurate and concise responses compared with other baselines.

Comparable runtime. We report the runtime of different algorithms in Table 2. With one online sample per question ($G = 1$), sampling adds little overhead, especially for the larger model. In contrast, $G = 5$ roughly doubles the cost for only a slight performance improvement; see Table 1. Therefore, $G = 1$ is an ideal choice.

5.2 SAFETY-AWARE FINE-TUNING

For safety-aware LLM fine-tuning task, we follow the setup from (Shen et al., 2025a). We use REDORCA datasets as the lower-level raw dataset, which mixes SLIMORCA dataset with 22k potentially unsafe instructions and responses from the ANTHROPIC RED-TEAMING dataset (Ganguli et al., 2022). For the upper-level safety guidance, we use 1) the BLUEORCA dataset (Shen et al., 2025a) which contains safe data distilled from SLIMORCA and 2) `Dahoas/rm-hh-rlhf` dataset, a offline RL reward version of Anthropic's Helpful and Harmless preference dataset (Bai et al., 2022).

Dynamic masking further improves the safety. So far, the default masked question index set $\mathcal{I}_M$ is sampled once and held fixed in Algorithm 1. In safety aware fine-tuning, performance improves further by generating online responses only for questions currently flagged as unsafe, i.e., by dynamically constructing the following $\mathcal{I}_M$ for the bottom ranked questions

$$\textbf{Online dynamic:} \qquad \mathcal{I}_M^k = \{i \mid \sigma_i(\omega_t) \text{ is ranked bottom } R \text{ among all data}\},$$

where $R = \frac{N_M}{N}$ is the online sample ratio. We call this the online dynamic strategy. Table 3 shows that this dynamic strategy further improves performance on the unsafe dataset REDORCA, suggesting that question-level BDS scores and response-level BMO weights complement each other in practice.

Online refinement enhances the data quality. Online regeneration uses upper-level safe data to refine responses, even when the original responses are harmful. Table 3 shows improved validation performance and lower loss on the selected lower-level data, suggesting improved data quality. Online dynamic strategy also improves mean OOD QA accuracy over Offline BDS by 0.96% and reduces mean harmfulness from 1.203 to 1.143 and the score-5 rate from 4.3% to 3.0% relative to Offline BDS, as shown in Figure 3(d). These results complement validation loss and general QA accuracy; the harmfulness judgement details are in Appendix D.5.

6 CONCLUSIONS

We introduce a unified bilevel framework for offline question selection BDS and online self-refining generation. Theoretically, we establish when global BDS improves over direct mixing, and prove that gradient-based approach can converge to the optimal solution of BDS with one-layer Transformer architecture. To further improve the lower-level data quality, we propose online self-refining generation algorithm and show BDS–BMO equivalence, yielding closed-form online response weights. Experiments on LLM fine-tuning validate the effectiveness of the proposed online selection algorithm.

AI USE STATEMENT

We used generative AI assistants for manuscript polishing, discussion of theoretical arguments and their presentation, and plotting assistance. We also used GPT-4 Turbo as an LLM judge for ALPACAEVAL response-quality comparisons and GPT-4.1-mini for HEx-PHI harmfulness evaluation, as described in Section 5 and Appendix D.5. These automated judgments complement the reported held-out losses and zero-shot QA accuracy. The authors take responsibility for the paper's final content, including AI-assisted text, analyses, and claims.

ETHICS STATEMENT

This work studies bilevel data curation for improving the quality and safety of LLM supervised fine-tuning. Our experiments use publicly available datasets, including corpora containing harmful instructions and responses. Potential benefits include making better use of existing training data and reducing harmful behavior during fine-tuning. However, data selection can inherit biases and coverage gaps from the validation set, while online response regeneration may reproduce factual errors, stereotypes, or harmful content. Improvements on the evaluated datasets do not establish safety or robustness across all deployment settings, and LLM-based evaluations remain subject to judge bias and error. Deployment would require further evaluation, appropriate data safeguards, and safeguards against harmful outputs.

REPRODUCIBILITY STATEMENT

We present the bilevel formulations and offline and online update procedures in Sections 2–4, with the online algorithm summarized in Algorithm 1. The theoretical assumptions are stated alongside the results, with supporting derivations and proofs in Appendices A, B, and C. Section 5 describes the models, datasets, baselines, evaluation metrics, hardware, and repeated runs; Table 2 reports runtime comparisons. Appendix D.1 details the data sizes, length filtering, LoRA configuration, optimization hyperparameters, and response generation settings. Appendix D.2 provides qualitative output comparisons and analyses of the learned data-selection behavior. Appendices D.3–D.6 report per-benchmark four-run QA results, component-wise ablations, the harmfulness evaluation, and diagnostics of the practical approximations. The implementation builds on the BIPOST library referenced in Section 5.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proc. International Conference on Machine Learning*, pp. 242–252, 2019.

Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. In *Proc. International Conference on Learning Representations*, virtual, 2022.

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

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. Nested learning: The illusion of deep learning architectures. In *Proc. Advances in Neural Information Processing Systems*, San Diego, CA, 2025.

Omar Bennouna, Amine Bennouna, Saurabh Amin, and Asuman Ozdaglar. What data enables optimal decisions? an exact characterization for linear optimization. *arXiv preprint arXiv:2505.21692*, 2025.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *Proc. International Conference on Learning Representations*, Vienna, Austria, 2024.

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

Matthew M Botvinick, Yael Niv, and Andrew G Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Lisha Chen, Quan Xiao, Ellen Hidemi Fukuda, Xinyi Chen, Kun Yuan, and Tianyi Chen. Efficient first-order optimization on the pareto set for multi-objective learning under preference guidance. In *Proc. International Conference on Machine Learning*, Vancouver, BC, Canada, 2025.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Proc. Advances in Neural Information Processing Systems*, virtual, 2021.

Hyeong Kyu Choi, Xuefeng Du, and Yixuan Li. Safety-aware fine-tuning of large language models. In *Neurips Safe Generative AI Workshop*, 2024.

Kalyanmoy Deb and Ankur Sinha. Solving bilevel multi-objective optimization problems using evolutionary algorithms. In *International conference on evolutionary multi-criterion optimization*, pp. 110–124. Springer, 2009.

S. Dempe and P. Mehlitz. Semivectorial bilevel programming versus scalar bilevel programming. *Optimization*, 69(4):657–679, 2020.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proc. International Conference on Machine Learning*, Sydney, Australia, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proc. International Conference on Machine Learning*, Stockholm, Sweden, 2018.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL `https://zenodo.org/records/12608602`.

Hossein Shokri Ghadikolaei, Hadi Ghauch, Carlo Fischione, and Mikael Skoglund. Learning and data selection in big datasets. In *Proc. International Conference on Machine Learning*, 2019.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *Proc. International Conference on Machine Learning*, virtual, 2020.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim F Tekin, and Ling Liu. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. In *Proc. Advances in Neural Information Processing Systems*, volume 37, pp. 104521–104555, 2024a.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *Proc. International Conference on Machine Learning*, Vienna, Austria, 2024b.

Joshua Hang Sai Ip, Ankush Chakrabarty, Ali Mesbah, and Diego Romeres. User preference meets pareto-optimality in multi-objective bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20246–20254, 2025.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proc. International Conference on Machine Learning*, virtual, 2021.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Proc. Advances in Neural Information Processing Systems*, virtual, 2020.

Feiyang Kang, Hoang Anh Just, Yifan Sun, Himanshu Jahagirdar, Yuanzhi Zhang, Rongxing Du, Anit Kumar Sahu, and Ruoxi Jia. Get more for less: Principled data selection for warming up fine-tuning in llms. In *Proc. International Conference on Learning Representations*, Vienna, Austria, 2024.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In *Proc. Advances in Neural Information Processing Systems*, virtual, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations*, 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20 (7):512–534, 2016.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *Proc. International Conference on Machine Learning*, Honolulu, HI, 2023.

Junyi Li, Bin Gu, and Heng Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proc. AAAI Conference on Artificial Intelligence*, virtual, 2022.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, May 2023.

Zeman Li, Yuan Deng, Peilin Zhong, Meisam Razaviyayn, and Vahab Mirrokni. PiKE: Adaptive data mixing for multi-task learning under low gradient conflicts. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025.

Zihao Li, Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason Klusowski, Jianqing Fan, and Mengdi Wang. One-layer transformer provably learns one-nearest neighbor in context. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2024.

Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. Data-efficient fine-tuning for LLM-based recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 365–374, 2024.

Yong Lin, Shange Tang, Bohan Lyu, Ziran Yang, Jui-Hui Chung, Haoyu Zhao, Lai Jiang, Yihan Geng, Jiawei Ge, Jingruo Sun, et al. Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. *arXiv preprint arXiv:2508.03613*, 2025.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *Proc. International Conference on Learning Representations*, Singapore, Singapore, 2025.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *Proc. International Conference on Machine Learning*, virtual, 2020.

Risheng Liu, Yaohua Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *Proc. International Conference on Machine Learning*, Honolulu, HI, 2023.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *Proc. International Conference on Machine Learning*, Honolulu, HI, 2023.

Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, et al. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*, 2025.

Debabrata Mahapatra and Vaibhav Rajan. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *Proc. International Conference on Machine Learning*, virtual, 2020.

Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and Sueyeon Chung. Emergence of separable manifolds in deep language representations. In *Proc. International Conference on Machine Learning*, virtual, 2020.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proc. International Conference on Machine Learning*, New York City, NY, 2016.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proc. International Conference on Learning Representations*, 2024.

Yi Ren and Danica J Sutherland. Learning dynamics of LLM Finetuning. In *Proc. International Conference on Learning Representations*, Singapore, Singapore, 2025.

Yinuo Ren, Tesi Xiao, Michael Shavlovsky, Lexing Ying, and Holakou Rahmanian. COS-DPO: Conditioned one-shot multi-objective fine-tuning framework. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Baturay Saglam, Paul Kassianik, Blaine Nelson, Sajana Weerawardhena, Yaron Singer, and Amin Karbasi. Large language models encode semantics in low-dimensional linear subspaces. *arXiv preprint arXiv:2507.09709*, 2025.

Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. SEAL: Safety-enhanced aligned llm fine-tuning via bilevel data selection. In *Proc. International Conference on Learning Representations*, 2025a.

Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method. *Mathematical Programming*, 214:539–589, 2025b.

Bingqing Song, Boran Han, Shuai Zhang, Jie Ding, and Mingyi Hong. Unraveling the gradient descent dynamics of transformers. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2024a.

Yuda Song, Gokul Swamy, Aarti Singh, J Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2024b.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage suboptimal, on-policy data. In *Proc. International Conference on Machine Learning*, Vienna, Austria, 2024.

Hiroki Tanabe, Ellen H Fukuda, and Nobuo Yamashita. New merit functions for multiobjective optimization and their properties. *Optimization*, 73(13):3821–3858, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

Dorothy Tse, Rosamund F Langston, Masaki Kakeyama, Ingrid Bethus, Patrick A Spooner, Emma R Wood, Menno P Witter, and Richard GM Morris. Schemas and memory consolidation. *Science*, 316(5821):76–82, 2007.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems*, 2017.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024a.

Jiachen T Wang, Tong Wu, Dawn Song, Prateek Mittal, and Ruoxi Jia. Greats: Online selection of high-quality data for llm training in every iteration. In *Proc. Advances in Neural Information Processing Systems*, 2024b.

Jiali Wang, He Chen, Rujun Jiang, Xudong Li, and Zihao Li. Fast algorithms for stackelberg prediction game with least squares loss. In *Proc. International Conference on Machine Learning*, virtual, 2021.

Jiali Wang, Wen Huang, Rujun Jiang, Xudong Li, and Alex L Wang. Solving stackelberg prediction game with least squares loss via spherically constrained least squares reformulation. In *Proc. International Conference on Machine Learning*, Baltimore, MD, 2022.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proc. of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *Proc. International Conference on Learning Representations*, 2022.

Yongtao Wu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. On the convergence of encoder-only shallow transformers. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2023.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *Proc. International Conference on Machine Learning*, Vienna, Austria, 2024.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Peiyao Xiao, Chaosheng Dong, Shaofeng Zou, and Kaiyi Ji. Ldc-mtl: Balancing multi-task learning through scalable loss discrepancy control. *arXiv preprint arXiv:2502.08585*, 2025.

Quan Xiao and Tianyi Chen. Unlocking global optimality in bilevel optimization: A pilot study. In *Proc. International Conference on Learning Representations*, Singapore, Singapore, 2025.

Quan Xiao, Songtao Lu, and Tianyi Chen. A generalized alternating method for bilevel optimization under the polyak-łojasiewicz condition. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2023.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2023a.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2023b.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. In *Proc. Advances in Neural Information Processing Systems*, virtual, 2021.

Runtian Zhai, Chen Dan, J Zico Kolter, and Pradeep Kumar Ravikumar. Understanding why generalized reweighting does not improve over ERM. In *Proc. International Conference on Learning Representations*, Kigali, Rwanda, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Shuai Zhang, Meng Weng, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis. In *Proc. International Conference on Learning Representations*, virtual, 2022.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *arXiv preprint arXiv:2508.11408*, 2025.

Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. An introduction to bi-level optimization: Foundations and applications in signal processing and machine learning. *arXiv preprint arXiv:2308.00788*, 2023.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto alignment via preference adaptation for llms. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2023.

Supplementary Material

Contents

A Convexity of SFT loss with respect to the backbone model **17**

B Proofs of the main theorems **19**

B.1 Proof of Theorem 1 19

B.2 Proof of Proposition 1 20

B.3 Proof of Theorem 2 20

B.4 Proof of Theorem 3 36

C Implicit weight assigned by `BMO` **36**

D Additional experiments **38**

D.1 Details of experimental setup 38

D.2 Quality comparisons of model outputs 39

D.3 Per-benchmark OOD QA evaluation 43

D.4 Component-wise ablation 44

D.5 Harmfulness evaluation 44

D.6 Importance-ratio approximated response-level weights 44

A Convexity of SFT loss with respect to the backbone model

With causal masking in attention (Vaswani et al., 2017), the backbone model $\phi_\theta(x, y)$ is unable to see the future tokens before predicting, even if we input the whole sequence (Ren & Sutherland, 2025). Therefore, the SFT loss can be viewed as the token-level cross-entropy loss of a sequential multi-class classification problem, where the label is the next token in the response y. Due to the nonlinearity of $\phi_\theta(x, y)$, SFT loss might not be convex with respect to θ, but it is convex with respect to the backbone representation $z = \phi_\theta(x, y) \in \mathbb{R}^{V\times D}$.

Lemma 6. *Per-sample SFT loss in (1) is convex with respect to the backbone model $z = \phi_\theta(x, y) \in \mathbb{R}^{V\times D}$.*

Due to the nonlinear backbone model $\phi_\theta(x, y)$, SFT loss might not be convex with respect to θ, which makes the optimization landscape complicated. Nevertheless, the building block of our theory lies in the convexity of the SFT loss with respect to the backbone representation $z = \phi_\theta(x, y) \in \mathbb{R}^{V\times D}$, which makes the SFT loss a composite convex function over θ.

Proof. According to (1), SFT loss takes the form of

$$\begin{aligned}\mathcal{L}_{\text{SFT}}(\theta; x, y) &= -\sum_{d=1}^{D} \mathbf{e}_{y_d}^\top \log \pi_\theta(y_d \mid x, y_{<d}) = -\sum_{d=1}^{D} \mathbf{e}_{y_d}^\top \log \sigma\left(\phi_\theta(x, y)\right)_{[:,d]} \\ &= -\sum_{d=1}^{D} \mathbf{e}_{y_d}^\top \log \sigma(z)_{[:,d]} = -\sum_{d=1}^{D} \mathbf{e}_{y_d}^\top \log \sigma(z_{[:,d]}). \end{aligned} \tag{11}$$

The Jacobian of softmax function is well-known (Gao & Pavel, 2017) and is given by

$$\nabla\sigma(z_{[:,d]}) = \operatorname{diag}\left(\sigma(z_{[:,d]})\right) - \sigma(z_{[:,d]})\sigma(z_{[:,d]})^\top.$$

Therefore, the derivative of $\mathcal{L}_{\text{SFT}}(\theta; x, y)$ with respect to $z_{[:,d]}$ is given by

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z_{[:,d]}} &= \nabla\sigma(z_{[:,d]}) \frac{\partial \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial \sigma(z_{[:,d]})} \\ &= -\left[\text{diag}\left(\sigma(z_{[:,d]})\right) - \sigma(z_{[:,d]})\sigma(z_{[:,d]})^\top\right] \left[\frac{(\mathbf{e}_{y_d})_1}{\sigma(z_{[:,d]})_1}, \cdots, \frac{(\mathbf{e}_{y_d})_V}{\sigma(z_{[:,d]})_V}\right]^\top \\ &\overset{(a)}{=} \sigma(z_{[:,d]}) - \mathbf{e}_{y_d} \end{aligned} \tag{12}$$

where $(\mathbf{e}_{y_d})_i$ denotes the i-the element in the one-hot vector $\mathbf{e}_{y_d} \in \mathbb{R}^V$ and $\sigma(z_{[:,d]})_i$ denotes the i-the element in the softmax vector $\sigma(z_{[:,d]}) \in \mathbb{R}^V$, and (a) uses the fact that $\sum_i (\mathbf{e}_{y_d})_i = 1$. Moreover, the Jacobian of $\mathcal{L}_{\text{SFT}}(\theta; x, y)$ with respect to z is given by

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z} &= \left[\frac{\partial \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z_{[:,1]}}, \cdots, \frac{\partial \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z_{[:,D]}}\right] \\ &= \left[\sigma(z_{[:,1]}) - \mathbf{e}_{y_1}, \cdots, \sigma(z_{[:,D]}) - \mathbf{e}_{y_D}\right] \end{aligned} \tag{13}$$

On the other hand, the second-order derivative of $\mathcal{L}_{\text{SFT}}(\theta; x, y)$ with respect to $z_{[:,d]}$ is given by

$$\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial^2 z_{[:,d]}} = \nabla\sigma(z_{[:,d]}), \text{ and } \frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z_{[:,d]} \partial z_{[:,\tilde{d}]}} = \mathbf{0}^{V\times D}, \text{ if } \tilde{d} \neq d. \tag{14}$$

where $\mathbf{0}^{V\times D} \in \mathbb{R}^{V\times D}$ is the zero matrix. The Hessian of $\mathcal{L}_{\text{SFT}}(\theta; x, y)$ over z is defined as the Hessian of the vectorized z, i.e. $\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z} \in \mathbb{R}^{Vd\times Vd}$. We define the vectorized z as

$$\text{vec}(z) = [z_{11}, z_{21}, \cdots, z_{V1}, \cdots, z_{1D}, z_{2n}, \cdots z_{VD}] \tag{15}$$

Let $i = (d_1 - 1)V + v_1$ and $j = (d_2 - 1)V + v_2$, then each element of the Hessian is

$$\left[\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z}\right]_{ij} = \left[\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial \,\text{vec}(z)}\right]_{ij} = \begin{cases} \nabla\sigma(z_{[:,d_1]})_{v_1,v_2} & \text{if } d_1 = d_2 \\ 0 & \text{if } d_1 \neq d_2 \end{cases} \tag{16}$$

Then for any matrix $u \in \mathbb{R}^{V\times D}$, it holds that

$$\begin{aligned}&\text{vec}(u)^\top \frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z} \text{vec}(u) \\ &= \sum_{i=1}^{VD}\sum_{j=1}^{VD} \text{vec}(u)_i^\top \left[\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z}\right]_{ij} \text{vec}(u)_j \\ &= \sum_{d_1=1}^{D}\sum_{d_2=1}^{D}\sum_{v_1=1}^{V}\sum_{v_2=1}^{V} \text{vec}(u)_{((d_1-1)V+v_1)} \left[\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z}\right]_{((d_1-1)V+v_1)((d_2-1)V+v_2)} \text{vec}(u)_{((d_2-1)V+v_2)} \\ &\overset{(16)}{=} \sum_{d=1}^{D}\sum_{v_1=1}^{V}\sum_{v_2=1}^{V} \text{vec}(u)_{((d-1)V+v_1)} \left[\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z}\right]_{((d-1)V+v_1)((d-1)V+v_2)} \text{vec}(u)_{((d-1)V+v_2)} \\ &= \sum_{d=1}^{D}\sum_{v_1=1}^{V}\sum_{v_2=1}^{V} u_{v_1,d} \left[\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z}\right]_{((d-1)V+v_1)((d-1)V+v_2)} u_{v_2,d} \\ &\overset{(16)}{=} \sum_{d=1}^{D}\sum_{v_1=1}^{V}\sum_{v_2=1}^{V} u_{v_1,d} \nabla\sigma(z_{[:,d]})_{v_1,v_2} u_{v_2,d} \\ &= \sum_{d=1}^{D} u_{[:,d]}^\top \nabla\sigma(z_{[:,d]}) u_{[:,d]} \geq 0 \end{aligned}$$

where the last inequality follows from the positive semi-definiteness of each gradient of softmax function $\nabla\sigma(z_{[:,d]})$ (Gao & Pavel, 2017). Therefore, the Hessian $\frac{\partial^2 \mathcal{L}_{\text{SFT}}(\theta; x, y)}{\partial z}$ is positive semi-definite for any z, suggesting that $\mathcal{L}_{\text{SFT}}(\theta; x, y)$ is convex over z, which completes the proof.

□

B PROOFS OF THE MAIN THEOREMS

B.1 PROOF OF THEOREM 1

Let $v_* := \min_\theta \mathcal{L}_{\mathrm{val}}(\theta)$ and assume there exists a feasible pair $(\bar{\omega}, \bar{\theta})$ for BDS such that $\mathcal{L}_{\mathrm{val}}(\bar{\theta}) = v_*$. Because $\min \mathcal{L}_{\mathrm{val}}$ and $\min \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i)$ do not share the optimal model θ, then we know

$$\Delta_{\mathrm{val}} := \min_{\theta \in \mathcal{S}_{\mathrm{val}}} \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i) > 0,$$

$$\Delta_{\mathrm{sft}} := \min_{\theta \in \mathcal{S}_{\mathrm{sft}}} \mathcal{L}_{\mathrm{val}}(\theta) - v_* > 0.$$

We will show that the conclusion of Theorem 1 holds with the explicit threshold

$$\rho_c := \frac{\Delta_{\mathrm{sft}}}{\Delta_{\mathrm{val}} + \Delta_{\mathrm{sft}}}.$$

Proof. By assumption, there exists a feasible pair $(\bar{\omega}, \bar{\theta})$ for BDS such that $\mathcal{L}_{\mathrm{val}}(\bar{\theta}) = v_*$. Since BDS minimizes the validation loss over its feasible set, any global solution (ω^*, θ^*) satisfies

$$\mathcal{L}_{\mathrm{val}}(\theta^*) \le \mathcal{L}_{\mathrm{val}}(\bar{\theta}) = v_*.$$

Because $v_* = \min_\theta \mathcal{L}_{\mathrm{val}}(\theta)$ is the global minimum of the validation objective, we also have $\mathcal{L}_{\mathrm{val}}(\theta^*) \ge v_*$. Hence $\mathcal{L}_{\mathrm{val}}(\theta^*) = v_*$.

Fix any $\rho > \rho_c$. Let $\tilde{\theta} \in \mathcal{S}_{\mathrm{mix}}(\rho)$ be a global minimizer of the direct-mixing objective

$$\frac{\rho}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i) + (1-\rho)\mathcal{L}_{\mathrm{val}}(\theta).$$

We claim that $\mathcal{L}_{\mathrm{val}}(\tilde{\theta}) > v_*$. Otherwise, we have $\mathcal{L}_{\mathrm{val}}(\tilde{\theta}) = v_*$. Then, by the definition of Δ_{val},

$$\frac{\rho}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\tilde{\theta}; x^i, y^i) + (1-\rho)\mathcal{L}_{\mathrm{val}}(\tilde{\theta}) \ge \rho\Delta_{\mathrm{val}} + (1-\rho)v_*.$$

On the other hand, choose any $\theta_{\mathrm{sft}} \in \mathcal{S}_{\mathrm{sft}}$ such that

$$\mathcal{L}_{\mathrm{val}}(\theta_{\mathrm{sft}}) = v_* + \Delta_{\mathrm{sft}},$$

which exists by the definition of Δ_{sft}. Since $\frac{1}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta_{\mathrm{sft}}; x^i, y^i) = 0$ under Assumption 1, its direct-mixing objective value is

$$\frac{\rho}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta_{\mathrm{sft}}; x^i, y^i) + (1-\rho)\mathcal{L}_{\mathrm{val}}(\theta_{\mathrm{sft}}) = (1-\rho)(v_* + \Delta_{\mathrm{sft}}).$$

Because

$$\rho > \rho_c = \frac{\Delta_{\mathrm{sft}}}{\Delta_{\mathrm{val}} + \Delta_{\mathrm{sft}}} \quad \Longleftrightarrow \quad \rho\Delta_{\mathrm{val}} > (1-\rho)\Delta_{\mathrm{sft}},$$

we obtain

$$\begin{aligned}\frac{\rho}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\tilde{\theta}; x^i, y^i) + (1-\rho)\mathcal{L}_{\mathrm{val}}(\tilde{\theta}) &= \rho\Delta_{\mathrm{val}} + (1-\rho)v_* > (1-\rho)(v_* + \Delta_{\mathrm{sft}}) \\ &= \frac{\rho}{N}\sum_{i=1}^N \mathcal{L}_{\mathrm{SFT}}(\theta_{\mathrm{sft}}; x^i, y^i) + (1-\rho)\mathcal{L}_{\mathrm{val}}(\theta_{\mathrm{sft}}),\end{aligned}$$

which contradicts the optimality of $\tilde{\theta}$. Therefore, every direct-mixing minimizer satisfies $\mathcal{L}_{\mathrm{val}}(\tilde{\theta}) > v_*$.

Combining this with $\mathcal{L}_{\mathrm{val}}(\theta^*) = v_*$ yields

$$\mathcal{L}_{\mathrm{val}}(\theta^*) = v_* < \mathcal{L}_{\mathrm{val}}(\tilde{\theta}),$$

which completes the proof. □

Algorithm 2 PBGD for offline BDS

1: Initialize ω^0 and $\theta^0 = (W_{kq}^0, W_{ov}^0)$; choose step sizes $\{\eta_k\}_{k\geq 0}$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Update all variables by the full-gradient penalty step
4:

$$\omega^{k+1} = \omega^k - \eta_k \nabla_\omega \mathcal{L}_\gamma(\omega^k, \theta^k), \tag{17a}$$

$$W_{ov}^{k+1} = W_{ov}^k - \eta_k \nabla_{W_{ov}} \mathcal{L}_\gamma(\omega^k, \theta^k), \tag{17b}$$

$$W_{kq}^{k+1} = W_{kq}^k - \eta_k \nabla_{W_{kq}} \mathcal{L}_\gamma(\omega^k, \theta^k). \tag{17c}$$

5: **end for**

B.2 Proof of Proposition 1

Proof. Let (ω^*, θ^*) be any global solution of BDS, and write $\lambda^* := \sigma(\omega^*) \in \Delta^N$. Let

$$\mathcal{U} := \left\{ i \in [N] \;\middle|\; \exists \theta \in \mathcal{S}_{\mathrm{val}} \text{ such that } \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i) = 0 \right\}.$$

We first show that $\theta^* \in \mathcal{S}_{\mathrm{val}}$. By Assumption 2, there exists an index $i_0 \in [N]$ such that

$$\min_{\theta \in \mathcal{S}_{\mathrm{val}}} \mathcal{L}_{\mathrm{SFT}}(\theta; x^{i_0}, y^{i_0}) = 0.$$

Choose $\bar{\theta} \in \mathcal{S}_{\mathrm{val}}$ satisfying $\mathcal{L}_{\mathrm{SFT}}(\bar{\theta}; x^{i_0}, y^{i_0}) = 0$, and let $\bar{\lambda} = e_{i_0} \in \Delta^N$. Then $\bar{\theta} \in \arg\min_{\theta'} \frac{1}{N} \sum_{i=1}^N \bar{\lambda}_i \mathcal{L}_{\mathrm{SFT}}(\theta'; x^i, y^i)$, so $(\bar{\lambda}, \bar{\theta})$ is feasible for BDS. Since $\mathcal{L}_{\mathrm{val}}(\bar{\theta}) = v_*$, every global solution of BDS must satisfy

$$\mathcal{L}_{\mathrm{val}}(\theta^*) = v_*,$$

that is, $\theta^* \in \mathcal{S}_{\mathrm{val}}$.

Next, because (λ^*, θ^*) is feasible for BDS and Assumption 1 gives a parameter $\hat{\theta}$ with $\mathcal{L}_{\mathrm{SFT}}(\hat{\theta}; x^i, y^i) = 0$ for all $i \in [N]$, the optimal value of the lower-level weighted problem is 0. Hence

$$\frac{1}{N} \sum_{i=1}^N \lambda_i^* \mathcal{L}_{\mathrm{SFT}}(\theta^*; x^i, y^i) = 0.$$

Since every term is nonnegative, $\lambda_i^* > 0$ implies

$$\mathcal{L}_{\mathrm{SFT}}(\theta^*; x^i, y^i) = 0.$$

Because $\theta^* \in \mathcal{S}_{\mathrm{val}}$, every index in the support of λ^* belongs to the useful set $\mathcal{U}$. Therefore

$$\sum_{i \in \mathcal{U}} \lambda_i^* = \sum_{i=1}^N \lambda_i^* = 1,$$

which proves the first claim.

For direct mixing, the lower-level weights are always uniform, namely

$$\lambda^{\mathrm{mix}} := \left(\frac{1}{N}, \ldots, \frac{1}{N} \right).$$

Therefore

$$\sum_{i \notin \mathcal{U}} \lambda_i^{\mathrm{mix}} = \sum_{i \notin \mathcal{U}} \frac{1}{N} = 1 - \frac{|\mathcal{U}|}{N}.$$

If useless samples exist, then $|\mathcal{U}| < N$, so $\sum_{i \notin \mathcal{U}} \lambda_i^{\mathrm{mix}} > 0$ for every $\rho > 0$. □

B.3 Proof of Theorem 2

In this section, we provide the proof of axillary lemmas and theorems for proving Theorem 2.

First, the deterministic offline PBGD iterates are summarized in Algorithm 2. We then introduce additional notations in this section. Let a denote the embedding dimension. For each training sample

(x^i, y^i) and token position $d \in [D]$, let $X_d^i = \mathsf{Rep}(x^i, y^i_{<d}) \in \mathbb{R}^{a \times L_d}$ denote the masked context representation used to predict y_d^i, where $\mathsf{Rep}(\cdot)$ denotes the embedding and causal-masking operation and $L_d = d_x + d - 1$. Let $X^i_{d,-1} \in \mathbb{R}^a$ denote the query column in X_d^i associated with the prediction of y_d^i. Similarly, for each validation sample $(\tilde{x}^j, \tilde{y}^j)$, define $\tilde{X}_d^j = \mathsf{Rep}(\tilde{x}^j, \tilde{y}^j_{<d}) \in \mathbb{R}^{a \times L_d}$ and $\tilde{X}^j_{d,-1} \in \mathbb{R}^a$. We define

$$\mathcal{S}_{\text{token}} = \{(X_d^i, y_d^i)\}_{i\in[N], d\in[D]} \cup \{(\widetilde{X}_d^j, \widetilde{y}_d^j)\}_{j\in[N'], d\in[D]}.$$

For any token-level training or validation sample$(X, y) \in \mathcal{S}_{\text{token}}$, in the one-layer softmax Transformer, define

$$B_\theta(X) := X^\top W_{kq} X_{-1}, \quad h_\theta(X) := X\sigma(B_\theta(X)), \quad U_\theta(X) := W_{ov} h_\theta(X),$$

and the weighted token-level feature matrix

$$H(\omega, \theta) := \left[\left\{ \frac{1}{\sqrt{N'D}} h_\theta(\widetilde{X}_d^j) \right\}_{j\in[N'],\, d\in[D]}, \left\{ \sqrt{\frac{\gamma\sigma_i(\omega)}{ND}} h_\theta(X_d^i) \right\}_{i\in[N],\, d\in[D]} \right] \in \mathbb{R}^{a\times(N'+N)D}.$$

Similarly, define the weighted logit-gradient matrix

$$G(\omega, \theta) := \left[\left\{ \frac{1}{\sqrt{N'D}} \nabla g_{\widetilde{y}_d^j}\Big(U_\theta(\widetilde{X}_d^j)\Big) \right\}_{j\in[N'],\, d\in[D]}, \left\{ \sqrt{\frac{\gamma\sigma_i(\omega)}{ND}} \nabla g_{y_d^i}\Big(U_\theta(X_d^i)\Big) \right\}_{i\in[N],\, d\in[D]} \right] \in \mathbb{R}^{V\times(N'+N)D}$$

where the columns of $G(\omega, \theta)$ and $H(\omega, \theta)$ are ordered in the same way.

For notational simplicity, along the offline PBGD trajectory we write $H^k := H(\omega^k, \theta^k)$. We assume the overparameterized regime $a \geq (N+N')D$, so that the feature matrix $H(\omega, W_{kq}) \in \mathbb{R}^{a\times(N+N')D}$ can be full column rank, i.e., the condition $\sigma_{\min}(H(\omega, W_{kq})) > 0$.

For a single token y, define $g_y(u) := -\mathbf{e}_y^\top \log\sigma(u) = -\log[\sigma(u)]_y$. For any $(X, y) \in \mathcal{S}_{\text{token}}$, define $l(\theta; X, y) = g_y(U_\theta(X))$. Then

$$\mathcal{L}_{\text{SFT}}(\theta; x^i, y^i) = \frac{1}{D}\sum_{d=1}^D g_{y_d^i}\big(U_\theta(X_d^i)\big) = \frac{1}{D}\sum_{d=1}^D l(\theta; X_d^i, y_d^i),$$

$$\mathcal{L}_{\text{SFT}}(\theta; \tilde{x}^j, \tilde{y}^j) = \frac{1}{D}\sum_{d=1}^D g_{\tilde{y}_d^j}\Big(U_\theta(\tilde{X}_d^j)\Big) = \frac{1}{D}\sum_{d=1}^D l(\theta; \tilde{X}_d^j, \tilde{y}_d^j).$$

For any integer $m \geq 2$, we use σ to denote the softmax map on $\mathbb{R}^m$. Its Jacobian is denoted by

$$J_\sigma(z) := \text{Diag}(\sigma(z)) - \sigma(z)\sigma(z)^\top \in \mathbb{R}^{m\times m}, \qquad z \in \mathbb{R}^m.$$

Assumption 3 (Bounded token-level geometry)**.** *For the tokenized input of training samples X^i and validation samples $\tilde{X}^j$, define*

$$B_X := \max\left\{ \max_{i\in[N],\, d\in[D]} \sigma_{\max}(X_d^i), \max_{j\in[N'],\, d\in[D]} \sigma_{\max}(\tilde{X}_d^j) \right\},$$

and

$$B_{-1} := \max\left\{ \max_{i\in[N],\, d\in[D]} \|X^i_{d,-1}\|_2, \max_{j\in[N'],\, d\in[D]} \|\tilde{X}^j_{d,-1}\|_2 \right\}.$$

We assume that $B_X < \infty$ and $B_{-1} < \infty$.

B.3.1 Sample-level regularity

We denote two parameters $\theta = (W_{ov}, W_{kq})$, $\theta' = (W'_{ov}, W'_{kq})$. We also denote $R(\theta, \theta') := \max\{\|W_{ov}\|_F, \|W'_{ov}\|_F\}$. Under Assumption 3, all token-level samples considered in this subsection satisfy $\sigma_{\max}(X) \leq B_X$, $\|X_{-1}\|_2 \leq B_{-1}$.

Lemma 7 (Sample gradient formulas)**.** *Under the above notations, for any $(X, y) \in \mathcal{S}_{\text{token}}$, the sample gradients satisfy*

$$\nabla_{W_{\text{ov}}} l(\theta; X, y) = \big(\sigma(U_\theta(X)) - \mathbf{e}_{I(y)}\big) h_\theta(X)^\top,$$

$$\nabla_{W_{\mathrm{kq}}} l(\theta; X, y) = X J_\sigma(B_\theta(X)) X^\top W_{\mathrm{ov}}^\top \big(\sigma(U_\theta(X)) - \mathbf{e}_{I(y)}\big) X_{-1}^\top.$$

Lemma 8 (Gradient formulas for the penalty objective)**.** *Under the above notations, recall that*

$$\mathcal{L}_\gamma(\omega, \theta) = \frac{1}{N'} \sum_{j=1}^{N'} \mathcal{L}_{\mathrm{SFT}}(\theta; \widetilde{X}^{\,j}, \widetilde{y}^{\,j}) + \gamma \sum_{i=1}^{N} [\sigma(\omega)]_i \, \mathcal{L}_{\mathrm{SFT}}(\theta; X^i, y^i).$$

Then the gradients with respect to $W_{\mathrm{ov}}, W_{\mathrm{kq}}$ and ω satisfy

$$\nabla_{W_{\mathrm{ov}}} \mathcal{L}_\gamma(\omega, \theta) = \frac{1}{N'D} \sum_{j=1}^{N'} \sum_{d=1}^{D} \nabla_{W_{\mathrm{ov}}} l(\theta; \widetilde{X}_d^{\,j}, \widetilde{y}_d^{\,j}) + \frac{\gamma}{ND} \sum_{i=1}^{N} \sum_{d=1}^{D} [\sigma(\omega)]_i \, \nabla_{W_{\mathrm{ov}}} l(\theta; X_d^i, y_d^i),$$

$$\nabla_{W_{\mathrm{kq}}} \mathcal{L}_\gamma(\omega, \theta) = \frac{1}{N'D} \sum_{j=1}^{N'} \sum_{d=1}^{D} \nabla_{W_{\mathrm{kq}}} l(\theta; \widetilde{X}_d^{\,j}, \widetilde{y}_d^{\,j}) + \frac{\gamma}{ND} \sum_{i=1}^{N} \sum_{d=1}^{D} [\sigma(\omega)]_i \, \nabla_{W_{\mathrm{kq}}} l(\theta; X_d^i, y_d^i),$$

and

$$\nabla_\omega \mathcal{L}_\gamma(\omega, \theta) = \frac{\gamma}{N} J_\sigma(\omega) \, \ell_{\mathrm{seq}}(\theta), \quad \ell_{\mathrm{seq}}(\theta) := \begin{bmatrix} \mathcal{L}_{\mathrm{SFT}}(\theta; X^1, y^1) \\ \vdots \\ \mathcal{L}_{\mathrm{SFT}}(\theta; X^N, y^N) \end{bmatrix} = \frac{1}{D} \begin{bmatrix} \sum_{d=1}^{D} l(\theta; X_d^1, y_d^1) \\ \vdots \\ \sum_{d=1}^{D} l(\theta; X_d^N, y_d^N) \end{bmatrix}.$$

Lemma 9 (Lipschitz continuity of softmax and its Jacobian)**.** *Let $m \geq 2$ be arbitrary, and let $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ be the softmax map. For any $z \in \mathbb{R}^m$, its Jacobian is*

$$J_\sigma(z) = \mathrm{Diag}(\sigma(z)) - \sigma(z)\sigma(z)^\top.$$

Then, for all $a, b, z \in \mathbb{R}^m$, the following hold:

(i) $\|J_\sigma(z)\|_2 \leq \frac{1}{2}, \qquad \|\sigma(a) - \sigma(b)\|_2 \leq \frac{1}{2}\|a - b\|_2.$

(ii) $\|J_\sigma(a) - J_\sigma(b)\|_2 \leq \frac{3}{2}\|a - b\|_2.$

(iii) $\|\sqrt{\sigma(a)} - \sqrt{\sigma(b)}\|_2 \leq \frac{\sqrt{2}}{4}\|a - b\|_2$, *where the square root is taken componentwise.*

Proof. For (i), let $p = \sigma(z)$. Direct differentiation gives

$$[J_\sigma(z)]_{ij} = p_i(\mathbf{1}_{\{i=j\}} - p_j), \qquad J_\sigma(z) = \mathrm{Diag}(p) - pp^\top.$$

For any $v \in \mathbb{R}^m$,

$$v^\top J_\sigma(z) v = \sum_{i=1}^{m} p_i v_i^2 - \left(\sum_{i=1}^{m} p_i v_i\right)^2 = \sum_{1 \leq i < j \leq m} p_i p_j (v_i - v_j)^2.$$

Using $(v_i - v_j)^2 \leq 2(v_i^2 + v_j^2)$ and $p_i(1 - p_i) \leq \frac{1}{2}$, we have

$$v^\top J_\sigma(z) v \leq 2 \sum_{i=1}^{m} p_i(1 - p_i) v_i^2 \leq \frac{1}{2}\|v\|_2^2.$$

Since $J_\sigma(z)$ is symmetric positive semidefinite, this implies $\|J_\sigma(z)\|_2 \leq 1/2$. And therefore $\|\sigma(a) - \sigma(b)\|_2 \leq \frac{1}{2}\|a - b\|_2$.

For (ii), let $p = \sigma(a)$, $q = \sigma(b)$, and $d = p - q$. Then

$$J_\sigma(a) - J_\sigma(b) = \mathrm{Diag}(d) - (pp^\top - qq^\top).$$

Since $pp^\top - qq^\top = pd^\top + dq^\top$, we obtain

$$\|J_\sigma(a) - J_\sigma(b)\|_2 \leq \|\,\mathrm{Diag}(d)\|_2 + \|pd^\top + dq^\top\|_2 \leq \|d\|_2 + (\|p\|_2 + \|q\|_2)\|d\|_2.$$

As p and q are probability vectors, $\|p\|_2 \leq 1$ and $\|q\|_2 \leq 1$. Hence

$$\|J_\sigma(a) - J_\sigma(b)\|_2 \leq 3\|\sigma(a) - \sigma(b)\|_2 \leq \frac{3}{2}\|a - b\|_2.$$

For (iii), define $F(z) := \sqrt{\sigma(z)}$, where the square root is componentwise. For $p = \sigma(z)$ and any direction $v \in \mathbb{R}^m$,

$$[DF(z)v]_i = \frac{1}{2\sqrt{p_i}}[J_\sigma(z)v]_i = \frac{1}{2}\sqrt{p_i}(v_i - p^\top v).$$

Therefore,

$$\|DF(z)v\|_2^2 = \frac{1}{4}\sum_{i=1}^{m} p_i(v_i - p^\top v)^2 = \frac{1}{4}v^\top J_\sigma(z)v \le \frac{1}{8}\|v\|_2^2.$$

Thus $\|DF(z)\|_2 \le \sqrt{2}/4$. Applying the fundamental theorem of calculus to F along the segment from b to a, we get

$$\|\sqrt{\sigma(a)} - \sqrt{\sigma(b)}\|_2 = \|F(a) - F(b)\|_2 \le \frac{\sqrt{2}}{4}\|a - b\|_2.$$

This completes the proof. □

Lemma 10 (Lipschitz continuity of the hidden feature map)**.** *Under Assumption 3. For any fixed sample X, the hidden feature map $h_\theta(X) = X\,\phi(X^\top W_{\mathrm{kq}} X_{-1})$ depends only on W_{kq} and satisfies*

$$\|h_{\theta'}(X) - h_\theta(X)\|_2 \le \frac{1}{2}\,B_X^2\,B_{-1}\,\|W'_{\mathrm{kq}} - W_{\mathrm{kq}}\|_F.$$
$$\|h_\theta(X)\|_2 \le B_X, \qquad \|h_{\theta'}(X)\|_2 \le B_X.$$

Proof. The lemma follows directly from Lemma 9 □

Lemma 11 (Blockwise Lipschitz continuity of the logit map)**.** *Under Assumption 3, for any fixed sample X, the softmax logit map $U_\theta(X) = W_{\mathrm{ov}} X \sigma(X^\top W_{\mathrm{kq}} X_{-1})$ depends on $\theta = (W_{\mathrm{ov}}, W_{\mathrm{kq}})$ and satisfies*

$$\|U_{\theta'}(X) - U_\theta(X)\|_2 \le \left(\sigma_X + \frac{1}{2}B_X^2 B_{-1} R(\theta,\theta')\right)\|\theta' - \theta\|_F. \tag{18}$$

where $R(\theta,\theta') = \min\{\sigma_{\max}(W_{\mathrm{ov}}), \sigma_{\max}(W'_{\mathrm{ov}})\}$

Proof. According to the definition of $U_\theta(X)$, we have

$$\begin{aligned}\|U_{\theta'}(X) - U_\theta(X)\|_2 &\le \|W'_{\mathrm{ov}}(h_{\theta'}(X) - h_\theta(X))\| + \|(W'_{\mathrm{ov}} - W_{\mathrm{ov}})h_\theta(X)\| \\ &\le \frac{1}{2}B_X^2\,B_{-1}\sigma_{\max}(W'_{\mathrm{ov}})\|W'_{kq} - W_{kq}\|_F + B_X\,\|W'_{\mathrm{ov}} - W_{\mathrm{ov}}\|_F \end{aligned} \tag{19}$$

Similarly, we can replace W'_{ov} with W_{ov} in (19). Therefore, we have

$$\begin{aligned}\|U_{\theta'}(X) - U_\theta(X)\|_2 &\le \frac{1}{2}B_X^2 B_{-1}\min\{\sigma_{\max}(W_{\mathrm{ov}}), \sigma_{\max}(W'_{\mathrm{ov}})\}\|W'_{\mathrm{kq}} - W_{\mathrm{kq}}\|_F + B_X\|W'_{\mathrm{ov}} - W_{\mathrm{ov}}\|_F \\ &\le \sqrt{B_X^2 + \frac{1}{4}B_X^4 B_{-1}^2\min\{\sigma_{\max}(W_{\mathrm{ov}}), \sigma_{\max}(W'_{\mathrm{ov}})\}^2}\|\theta' - \theta\|_F \\ &\le \left(B_X + \frac{1}{2}B_X^2 B_{-1}R(\theta,\theta')\right)\|\theta - \theta'\|_F.\end{aligned}$$

where $R(\theta,\theta') = \min\{\sigma_{\max}(W_{\mathrm{ov}}), \sigma_{\max}(W'_{\mathrm{ov}})\}$. □

Lemma 12 (Lipschitz continuity of the sequence SFT loss)**.** *Under Assumption 3, for any sequence-level sample (x, y) with token-level representations $\{(X_d, y_d)\}_{d=1}^D$, recall*

$$\mathcal{L}_{\mathrm{SFT}}(\theta; x, y) = \frac{1}{D}\sum_{d=1}^{D} l(\theta; X_d, y_d).$$

Then

$$|\mathcal{L}_{\mathrm{SFT}}(\theta'; x, y) - \mathcal{L}_{\mathrm{SFT}}(\theta; x, y)| \le \sqrt{2}\left(B_X + \frac{1}{2}B_X^2 B_{-1}R(\theta,\theta')\right)\|\theta' - \theta\|_F,$$

where $R(\theta,\theta') = \min\{\sigma_{\max}(W_{\mathrm{ov}}), \sigma_{\max}(W'_{\mathrm{ov}})\}$.

Proof. Recall that $l(\theta; X, y) := -\log\big(\mathbf{e}_{I(y)}^\top \sigma(U_\theta(X))\big) = g_y(U_\theta(X))$, since $\|\nabla g_y(u)\|_2 \le \sqrt{2}$, we have

$$|l(\theta'; X_d, y_d) - l(\theta; X_d, y_d)| \le \sqrt{2}\,\|U_{\theta'}(X_d) - U_\theta(X_d)\|_2.$$

Under Assumption 3, using the hidden-feature Lipschitz bound Lemma 11,

$$\|U_{\theta'}(X_d) - U_\theta(X_d)\|_2 \le \left(B_X + \frac{1}{2}B_X^2 B_{-1} R(\theta,\theta')\right)\|\theta' - \theta\|_F.$$

Therefore,

$$|l(\theta'; X_d, y_d) - l(\theta; X_d, y_d)| \le \sqrt{2}\left(B_X + \frac{1}{2}B_X^2 B_{-1} R(\theta,\theta')\right)\|\theta' - \theta\|_F.$$

Averaging over $d = 1, \dots, D$, we obtain

$$|\mathcal{L}_{\mathrm{SFT}}(\theta'; x, y) - \mathcal{L}_{\mathrm{SFT}}(\theta; x, y)| \le \sqrt{2}\left(B_X + \frac{1}{2}B_X^2 B_{-1} R(\theta,\theta')\right)\|\theta' - \theta\|_F.$$

□

Lemma 13 (Local Lipschitz continuity of the sample gradient)**.** *Under Assumption 3, for any sample $(X, y) \in \mathcal{S}_{\mathrm{token}}$ with token-level representations, recall its corresponding loss function $l(\theta; X, y) = -\log\left(\mathbf{e}_y^\top \sigma\left(U_\theta(X)\right)\right) = g_y(U_\theta(X))$, then*

$$\|\nabla_\theta l(\theta'; X, y) - \nabla_\theta l(\theta; X, y)\|_F \le 4B_X^2(1 + B_{-1})\left(1 + B_X B_{-1} R(\theta,\theta')\right)^2 \|\theta' - \theta\|_F,$$

where $R(\theta,\theta') := \min\{\sigma_{\max}(W_{\mathrm{ov}}), \sigma_{\max}(W'_{\mathrm{ov}})\}$.

Proof. By Lemma 7,

$$\nabla_{W_{\mathrm{ov}}} l(\theta; X, y) = \nabla g_y(U_\theta(X))\, h_\theta(X)^\top$$
$$\nabla_{W_{\mathrm{kq}}} l(\theta; X, y) = X\, J_\phi(B_\theta(X))\, X^\top W_{\mathrm{ov}}^\top \nabla g_y(U_\theta(X))\, X_{-1}^\top$$

We estimate the two blocks separately.

Step 1: estimate of the W_{ov}-block. By Assumption 3 and Lemma 11, we have

$$\begin{aligned}
&\|\nabla_{W_{\mathrm{ov}}} l(\theta'; X, y) - \nabla_{W_{\mathrm{ov}}} l(\theta; X, y)\|_F \\
&\le \|\nabla g_y(U_{\theta'}(X)) - \nabla g_y(U_\theta(X))\|_2 \, \|h_{\theta'}(X)\|_2 + \|\nabla g_y(U_\theta(X))\|_2 \, \|h_{\theta'}(X) - h_\theta(X)\|_2 \\
&\le \frac{1}{2}B_X \|U_{\theta'}(X) - U_\theta(X)\|_2 + \frac{\sqrt{2}}{2}B_X^2 B_{-1}\|W'_{\mathrm{kq}} - W_{\mathrm{kq}}\|_F \\
&\overset{(18)}{\le} \frac{1}{2}B_X\left(B_X + \frac{1}{2}B_X^2\, B_{-1} R(\theta,\theta')\right)\|\theta - \theta'\|_F + \frac{\sqrt{2}}{2}B_X^2 B_{-1}\|W'_{\mathrm{kq}} - W_{\mathrm{kq}}\|_F \\
&\le \left(\frac{1}{4}B_X^3 B_{-1} R(\theta,\theta') + \frac{1}{2}B_X^2(1 + \sqrt{2}B_{-1})\right)\|\theta - \theta'\|_F
\end{aligned} \tag{20}$$

Step 2: estimate of the W_{kq}-block. We denote

$$G_\theta(X) := X^\top W_{\mathrm{ov}}^\top \nabla g_y(U_\theta(X)).$$

Then we have

$$\begin{aligned}
&\|G_{\theta'}(X) - G_\theta(X)\|_2 \\
&\le \|X^\top W_{\mathrm{ov}}'^\top \left(\nabla g_y(U_{\theta'}(X)) - \nabla_y g(U_\theta(X))\right)\|_2 + \|X^\top \left(W'_{\mathrm{ov}} - W_{\mathrm{ov}}\right)^\top \nabla g_y(U_\theta(X))\| \\
&\le \frac{1}{2}B_X R(\theta,\theta')\|U_{\theta'}(X) - U_\theta(X)\|_2 + \sqrt{2}B_X\|W'_{\mathrm{ov}} - W_{\mathrm{ov}}\|_F \\
&\le \frac{1}{2}B_X R(\theta,\theta')\left(B_X + \frac{1}{2}B_X^2\, B_{-1} R(\theta,\theta')\right)\|\theta' - \theta\| + \sqrt{2}B_X\|W'_{\mathrm{ov}} - W_{\mathrm{ov}}\|_F
\end{aligned} \tag{21}$$

and

$$\|G_\theta(X)\|_2 \le \sqrt{2}B_X \sigma_{\max}(W_{\mathrm{ov}}) \tag{22}$$

Therefore, by Lemma 9, inequalities (22) and (22), we have

$$\begin{aligned}
&\|\nabla_{W_{\mathrm{kq}}} l(\theta'; X, y) - \nabla_{W_{\mathrm{kq}}} l(\theta; X, y)\|_F \\
&\le \|X\big(J_\phi(B_{\theta'}(X)) - J_\phi(B_\theta(X))\big)G_{\theta'}(X)X_{-1}^\top\| + \|X J_\phi(B_\theta(X))\big(G_{\theta'}(X) - G_\theta(X)\big)X_{-1}^\top\| \\
&\le B_X B_{-1}\Big(\|J_\phi(B_{\theta'}(X)) - J_\phi(B_\theta(X))\|_2 \, \|G_{\theta'}(X)\|_2 + \|J_\phi(B_\theta(X))\|_2 \, \|G_{\theta'}(X) - G_\theta(X)\|_2\Big)
\end{aligned}$$

$$\le \frac{3\sqrt{2}}{2}B_X^2 B_{-1} R(\theta,\theta')\|B_{\theta'}(X)-B_\theta(X)\|_2 + \frac{1}{2}B_X B_{-1}\|G_{\theta'}(X)-G_\theta(X)\|_2$$
$$\le \frac{3\sqrt{2}}{2}B_X^3 B_{-1}^2 R(\theta,\theta')\|W'_{\mathrm{kq}}-W_{\mathrm{kq}}\|_F + \frac{\sqrt{2}}{2}B_X^2 B_{-1}\|W'_{\mathrm{ov}}-W_{\mathrm{ov}}\|_F$$
$$+ \frac{1}{4}B_X^2 B_{-1} R(\theta,\theta')\left(B_X + \frac{1}{2}B_X^2\, B_{-1} R(\theta,\theta')\right)\|\theta'-\theta\|_F$$
$$\le \left(\frac{1}{8}B_X^4 B_{-1}^2 R(\theta,\theta')^2 + B_X^3 B_{-1}\left(\frac{3\sqrt{2}}{2}B_{-1}+\frac{1}{4}\right)R(\theta,\theta') + \frac{\sqrt{2}}{2}B_X^2 B_{-1}\right)\|\theta'-\theta\|_F \quad (23)$$

Step 3: combine the two blocks. Combining (20) and (23), we have

$$\|\nabla_\theta l(\theta';X,y)-\nabla_\theta l(\theta;X,y)\|_F$$
$$= \left(\|\nabla_{W_{\mathrm{ov}}} l(\theta';X,y)-\nabla_{W_{\mathrm{ov}}} l(\theta;X,y)\|_F^2 + \|\nabla_{W_{\mathrm{kq}}} l(\theta';X,y)-\nabla_{W_{\mathrm{kq}}} l(\theta;X,y)\|_F^2\right)^{\frac{1}{2}}$$
$$\le \|\nabla_{W_{\mathrm{ov}}} l(\theta';X,y)-\nabla_{W_{\mathrm{ov}}} l(\theta;X,y)\|_F + \|\nabla_{W_{\mathrm{kq}}} l(\theta';X,y)-\nabla_{W_{\mathrm{kq}}} l(\theta;X,y)\|_F$$
$$\le \left(\frac{1}{8}B_X^4 B_{-1}^2 R(\theta,\theta')^2 + B_X^3 B_{-1}\left(\frac{3\sqrt{2}}{2}B_{-1}+\frac{1}{2}\right)R(\theta,\theta') + B_X^2\left(\sqrt{2}B_{-1}+1\right)\right)\|\theta'-\theta\|_F$$
$$\le 4B_X^2(1+B_{-1})\left(1+B_X B_{-1} R(\theta,\theta')\right)^2\|\theta'-\theta\|_F$$

where $R(\theta,\theta') := \min\{\sigma_{\max}(W_{\mathrm{ov}}), \sigma_{\max}(W'_{\mathrm{ov}})\}$. □

Lemma 14 (Gradient norm bound for the sample loss)**.** *Under Assumption 3, for any sample $(X,y)\in\mathcal{S}_{\mathrm{token}}$ with token-level representation, then*

$$\|\nabla_\theta l(\theta;X,y)\|_F \le \sqrt{2}B_X + \frac{\sqrt{2}}{2}B_X^2 B_{-1}\sigma_{\max}(W_{\mathrm{ov}})$$

Proof. Fix one sample (X,y), and write

$$p := \sigma(U_\theta(X)), \qquad l(\theta;X,y) = -\log p_y.$$

By Lemma 7, we have

$$\nabla_{W_{\mathrm{ov}}} l(\theta;X,y) = \big(p-\mathbf{e}_y\big)h_\theta(X)^\top, \quad \nabla_{W_{\mathrm{kq}}} l(\theta;X,y) = X J_\sigma(X^\top W_{\mathrm{kq}} X_{-1}) X^\top W_{\mathrm{ov}}^\top \big(p-\mathbf{e}_y\big)X_{-1}^\top.$$

Under Assumption 3, by Lemma 9 and Lemma 10, we have

$$\|\nabla_{W_{\mathrm{ov}}} l(\theta;X,y)\|_F \le B_X\|p-\mathbf{e}_y\|_2 \le \sqrt{2}B_X,$$
$$\|\nabla_{W_{\mathrm{kq}}} l(\theta;X,y)\|_F \le \frac{1}{2}B_X^2 B_{-1}\sigma_{\max}(W_{\mathrm{ov}})\|p-\mathbf{e}_y\|_2 \le \frac{\sqrt{2}}{2}B_X^2 B_{-1}\sigma_{\max}(W_{\mathrm{ov}}).$$

Therefore,

$$\|\nabla_\theta l(\theta;X,y)\|_F \le \sqrt{2}B_X + \frac{\sqrt{2}}{2}B_X^2 B_{-1}\sigma_{\max}(W_{\mathrm{ov}})$$

□

Lemma 15 (Gradient norm bounds for the penalized SFT loss)**.** *Under Assumption 3, we have*

$$\|\nabla_{W_{\mathrm{ov}}}\mathcal{L}_\gamma(\omega,\theta)\|_F \le \sqrt{2}B_X\mathcal{L}_\gamma(\omega,\theta), \quad \|\nabla_{W_{\mathrm{kq}}}\mathcal{L}_\gamma(\omega,\theta)\|_F \le \frac{\sqrt{2}}{2}B_X^2 B_{-1}\sigma_{\max}(W_{\mathrm{ov}})\mathcal{L}_\gamma(\omega,\theta).$$

Consequently,

$$\|\nabla_\theta\mathcal{L}_\gamma(\omega,\theta)\|_F \le \sqrt{2}\left(B_X + \frac{1}{2}B_X^2 B_{-1}\sigma_{\max}(W_{\mathrm{ov}})\right)\mathcal{L}_\gamma(\omega,\theta).$$

Proof. Fix one sample (X,y), and write

$$p := \sigma(U_\theta(X)), \qquad l(\theta;X,y) = -\log p_y.$$

By Lemma 7, we have

$$\nabla_{W_{\mathrm{ov}}} l(\theta;X,y) = \big(p-\mathbf{e}_y\big)h_\theta(X)^\top, \quad \nabla_{W_{\mathrm{kq}}} l(\theta;X,y) = X J_\sigma(X^\top W_{\mathrm{kq}} X_{-1}) X^\top W_{\mathrm{ov}}^\top \big(p-\mathbf{e}_y\big)X_{-1}^\top.$$

Under Assumption 3, by Lemma 9 and Lemma 10, we have

$$\|\nabla_{W_{\rm ov}} l(\theta; X, y)\|_F \le B_X \|p - \mathbf{e}_y\|_2, \quad \|\nabla_{W_{\rm kq}} l(\theta; X, y)\|_F \le \frac{1}{2} B_X^2 B_{-1} \sigma_{\max}(W_{\rm ov}) \|p - \mathbf{e}_y\|_2.$$

Moreover,

$$\|p - \mathbf{e}_{I(y)}\|_2 = \left(\sum_{\ell \neq y} p_\ell^2 + (1 - p_y)^2 \right)^{1/2} \le \sqrt{2}(1 - p_y) \overset{(a)}{\le} \sqrt{2} l(\theta; X, y),$$

where we used $1 - t \le -\log t$ for $t \in (0, 1]$ in (a). Therefore,

$$\|\nabla_\theta l(\theta; X, y)\|_F \le \sqrt{2} \left(B_X + \frac{1}{2} B_X^2 B_{-1} \sigma_{\max}(W_{\rm ov}) \right) l(\theta; X, y)$$

By the triangle inequality,

$$\begin{aligned} \|\nabla_\theta \mathcal{L}_\gamma(\omega, \theta)\|_F &\le \frac{1}{N'D} \sum_{j=1}^{N'} \sum_{d=1}^{D} \|\nabla_\theta l(\theta; \widetilde{X}_d^j, \widetilde{y}_d^j)\|_F + \frac{\gamma}{ND} \sum_{i=1}^{N} \sum_{d=1}^{D} [\sigma(\omega)]_i \|\nabla_\theta l(\theta; X_d^i, y_d^i)\|_F \\ &\le \sqrt{2} \left(B_X + \frac{1}{2} B_X^2 B_{-1} \sigma_{\max}(W_{\rm ov}) \right) \mathcal{L}_\gamma(\omega, \theta). \end{aligned}$$

□

B.3.2 Proof of non-uniform smoothness condition

In this section, for any given point $z = (\omega, \theta)$ and stepsize η, we define

$$\begin{aligned} z(0) &= (\omega(0), \theta(0)) = (\omega, \theta) \\ z(s) &= (\omega(s), \theta(s)) = (\omega - s\eta \nabla_\omega L_\gamma(\omega, \theta), \theta - s\eta \nabla_\theta L_\gamma(\omega, \theta)). \end{aligned}$$

for any $s \in [0, 1]$.

Lemma 16 (Local smoothness of $\mathcal{L}_\gamma$)**.** *Under Assumption 3, for any $s \in [0, 1]$, we have*

$$\left\| \nabla_\theta \mathcal{L}_\gamma(\omega(s), \theta(s)) - \nabla_\theta \mathcal{L}_\gamma(\omega^k, \theta^k) \right\|_F \le L(\theta(0)) \|\theta(s) - \theta(0)\|_F$$

where

$$\begin{aligned} L(\theta(0)) := 4 \left(1 + \frac{\gamma}{N}\right) B_X^2 (1 + B_{-1}) \left(1 + B_X B_{-1} \sigma_{\max}(W_{\rm ov}(0))\right)^2, \\ + \frac{3\gamma}{\sqrt{N}} \left(\log V + B_X + B_X \left(1 + B_X B_{-1}\right) \sigma_{\max}(W_{\rm ov}(0)) \right). \end{aligned} \tag{24}$$

Proof. Denote $R(\theta(s), \theta(0)) = \min\{\sigma_{\max}(W_{\rm ov}(s)), \sigma_{\max}(W_{\rm ov}(0))\}$ for any $s \in [0, 1]$. Recall that

$$\mathcal{L}_\gamma(\omega, \theta) = \frac{1}{N'} \sum_{j=1}^{N'} \mathcal{L}_{\rm SFT}(\theta; \widetilde{X}^j, \widetilde{y}^j) + \frac{\gamma}{N} \sum_{i=1}^{N} [\phi(\omega)]_i \mathcal{L}_{\rm SFT}(\theta; X^i, y^i).$$

Step 1: Variation of the θ-gradient. We have

$$\begin{aligned} &\|\nabla_\theta \mathcal{L}_\gamma(\omega(s), \theta(s)) - \nabla_\theta \mathcal{L}_\gamma(\omega(0), \theta(0))\|_F \\ &\le \frac{1}{N'D} \sum_{j=1}^{N'} \sum_{d=1}^{D} \|\nabla_\theta l(\theta(s); \widetilde{X}_d^j, \widetilde{y}_d^j) - \nabla_\theta l(\theta(0); \widetilde{X}_d^j, \widetilde{y}_d^j)\|_F \\ &\quad + \frac{\gamma}{ND} \sum_{i=1}^{N} \sum_{d=1}^{D} [\sigma(\omega(s))]_i \|\nabla_\theta l(\theta(s); X_d^i, y_d^i) - \nabla_\theta l(\theta(0); X_d^i, y_d^i)\|_F \\ &\quad + \frac{\gamma}{ND} \sum_{i=1}^{N} \sum_{d=1}^{D} |\sigma(\omega(s))_i - \sigma(\omega(0))_i| \, \|\nabla_\theta l(\theta(0); X_i, y_i)\|_F \\ &\overset{(a)}{\le} 4 B_X^2 (1 + B_{-1}) \left(1 + B_X B_{-1} R(\theta, \theta')\right)^2 \left(1 + \frac{\gamma}{N}\right) \|\theta(s) - \theta(0)\|_F \end{aligned}$$

$$+\frac{\gamma}{ND}\|\sigma(\omega(s))-\sigma(\omega(0))\|_2\left(\sum_{i=1}^{N}\left(\sum_{d=1}^{D}\|\nabla_\theta l(\theta(0);X_d^i,y_d^i)\|_F\right)^2\right)^{\frac{1}{2}}$$

$$\overset{(b)}{\leq} 4B_X^2(1+B_{-1})\left(1+B_XB_{-1}R(\theta,\theta')\right)^2\left(1+\frac{\gamma}{N}\right)\|\theta(s)-\theta(0)\|_F$$

$$+\sqrt{\frac{2}{N}}\gamma\left(B_X+\frac{1}{2}B_X^2B_{-1}\sigma_{\max}(W_{\mathrm{ov}})\right)\|\sigma(\omega(s))-\sigma(\omega(0))\|_2 \tag{25}$$

where in (a) we use Lemma 13 and Cauchy Schwart inequality, in (b) we use Lemma 14.

Step 2: Variation of the ω-gradient.

By Lemma 8, we have

$$\nabla_\omega\mathcal{L}_\gamma(\omega,\theta)=\frac{\gamma}{N}J_\sigma(\omega)\ell_{\mathrm{seq}}(\theta),\qquad J_\sigma(\omega):=\mathrm{Diag}(\sigma(\omega))-\sigma(\omega)\sigma(\omega)^\top.$$

Therefore,

$$\begin{aligned}&\|\nabla_\omega\mathcal{L}_\gamma(\omega(s),\theta(s))-\nabla_\omega\mathcal{L}_\gamma(\omega(0),\theta(0))\|_2\\&\leq\frac{\gamma}{N}\|J_\sigma(\omega(s))-J_\sigma(\omega(0))\|_2\|\ell_{\mathrm{seq}}(\theta(0))\|_2+\frac{\gamma}{N}\|J_\sigma(\omega(s))\|_2\|\ell_{\mathrm{seq}}(\theta(s))-\ell_{\mathrm{seq}}(\theta(0))\|_2\\&\overset{(a)}{\leq}\frac{3\gamma}{2N}\|\ell_{\mathrm{seq}}(\theta(0))\|_2\|\omega(s)-\omega(0)\|_2+\frac{\gamma}{2N}\|\ell_{\mathrm{seq}}(\theta(s))-\ell_{\mathrm{seq}}(\theta(0))\|_2\end{aligned}$$

where in (a) we use 9. Next, by Lemma 12,

$$\begin{aligned}\|\ell_{\mathrm{seq}}(\theta(s))-\ell_{\mathrm{seq}}(\theta(0))\|_2&=\left(\sum_{i=1}^{N}|\mathcal{L}_{\mathrm{SFT}}(\theta(s);x_i,y_i)-\mathcal{L}_{\mathrm{SFT}}(\theta(0);x_i,y_i)|^2\right)^{1/2}\\&\leq\sqrt{2N}\left(B_X+\frac{1}{2}B_X^2B_{-1}R(\theta(s),\theta(0))\right)\|\theta(s)-\theta(0)\|_F.\end{aligned}$$

To bound $\|\ell(\theta(0))\|_2$, we use the uniform upper bound on each sample loss. Indeed, for every $i\in[N]$,

$$\begin{aligned}\mathcal{L}_{\mathrm{SFT}}(\theta(0);x_i,y_i)&\leq\mathcal{L}_{\mathrm{SFT}}(0;x_i,y_i)+|\mathcal{L}_{\mathrm{SFT}}(\theta(0);x_i,y_i)-\mathcal{L}_{\mathrm{SFT}}(0;x_i,y_i)|\\&\overset{(a)}{\leq}\log V+\sqrt{2}B_X\sigma_{\max}(W_{\mathrm{ov}}(0))\end{aligned}$$

where (a) also follows Lemma 12. Therefore, we have

$$\begin{aligned}\|\ell_{\mathrm{seq}}(\theta(0))\|_2&=\left(\sum_{i=1}^{N}\mathcal{L}_{\mathrm{SFT}}(\theta(0);x_i,y_i)^2\right)^{1/2}\\&\leq\sqrt{N}\left(\log V+\sqrt{2}B_X\sigma_{\max}(W_{\mathrm{ov}}(0))\right).\end{aligned}$$

Combining the above estimates, we obtain

$$\begin{aligned}&\|\nabla_\omega\mathcal{L}_\gamma(\omega(s),\theta(s))-\nabla_\omega\mathcal{L}_\gamma(\omega(0),\theta(0))\|_2\\&\leq\frac{3\gamma}{2\sqrt{N}}\left(\log V+\sqrt{2}B_X\sigma_{\max}(W_{\mathrm{ov}}(0))\right)\|\omega(s)-\omega(0)\|_2\\&\quad+\frac{\gamma}{2}\sqrt{\frac{2}{N}}\left(B_X+\frac{1}{2}B_X^2B_{-1}R(\theta(s),\theta(0))\right)\|\theta(s)-\theta(0)\|_F.\end{aligned}\tag{26}$$

Step 3: combine the two blocks. Combining (25) and (26), we have

$$\|\nabla\mathcal{L}_\gamma(\omega(s),\theta(s))-\nabla\mathcal{L}_\gamma(\omega(0),\theta(0))\|_F\leq L\|z(s)-z(0)\|_F.$$

$$\begin{aligned}L=&\,4B_X^2(1+B_{-1})\left(1+B_XB_{-1}R(\theta(s),\theta(0))\right)^2\left(1+\frac{\gamma}{N}\right)\\&+\frac{\gamma}{\sqrt{2N}}\left(B_X+\frac{1}{2}B_X^2B_{-1}\sigma_{\max}(W_{\mathrm{ov}}(0))\right)\end{aligned}$$

$$+\frac{3\gamma}{2\sqrt{N}}\left(\log V+\sqrt{2}B_X\sigma_{\max}(W_{\mathrm{ov}}(0))\right)$$
$$+\frac{\gamma}{2}\sqrt{\frac{2}{N}}\left(B_X+\frac{1}{2}B_X^2B_{-1}R(\theta(s),\theta(0))\right).$$

Since $R(\theta(s),\theta(0))\le\sigma_{\max}(W_{\mathrm{ov}}(0))$, we have

$$L\le 4B_X^2(1+B_{-1})\left(1+B_XB_{-1}\sigma_{\max}(W_{\mathrm{ov}}(0))\right)^2\left(1+\frac{\gamma}{N}\right)$$
$$+\frac{3\gamma}{\sqrt{N}}\left(\log V+B_X+B_X\left(1+B_XB_{-1}\right)\sigma_{\max}(W_{\mathrm{ov}}(0))\right)=L(\theta(0)).$$

Therefore,

$$\|\nabla\mathcal{L}_\gamma(\omega(s),\theta(s))-\nabla\mathcal{L}_\gamma(\omega(0),\theta(0))\|_F\le L(\theta(0))\|z(s)-z(0)\|_F.$$

where $L(\theta(0))$ is defined in (24). □

Lemma 17 (One-step Taylor upper bound from a one-sided gradient variation bound)**.** *For each iteration k, let $z_k(0)=(\omega^k,\theta^k)$ and*

$$z_k(s)=(\omega^k-s\eta_k\nabla_\omega\mathcal{L}_\gamma(\omega^k,\theta^k),\theta^k-s\eta_k\nabla_\theta\mathcal{L}_\gamma(\omega^k,\theta^k))$$

Then

$$\mathcal{L}_\gamma(z_k(1))\le\mathcal{L}_\gamma(z_k(0))+\langle\nabla\mathcal{L}_\gamma(z_k(0)),\,z_k(1)-z_k(0)\rangle+\frac{L_k}{2}\|z_k(1)-z_k(0)\|_F^2.$$

Equivalently,

$$\mathcal{L}_\gamma(\omega^{k+1},\theta^{k+1})\le\mathcal{L}_\gamma(\omega^k,\theta^k)+\langle\nabla\mathcal{L}_\gamma(\omega^k,\theta^k),\,z^{k+1}-z^k\rangle+\frac{L_k}{2}\|z^{k+1}-z^k\|_F^2.$$

where

$$L_k=4B_X^2(1+B_{-1})\left(1+B_XB_{-1}\sigma_{\max}(W_{\mathrm{ov}}^k)\right)^2\left(1+\frac{\gamma}{N}\right)$$
$$+\frac{3\gamma}{\sqrt{N}}\left(\log V+B_X+B_X\left(1+B_XB_{-1}\right)\sigma_{\max}(W_{\mathrm{ov}}^k)\right)\tag{27}$$

Proof. By the fundamental theorem of calculus,

$$\mathcal{L}_\gamma(z_k(1))-\mathcal{L}_\gamma(z_k(0))=\int_0^1\frac{d}{ds}\mathcal{L}_\gamma(z_k(s))\,ds=\int_0^1\langle\nabla\mathcal{L}_\gamma(z_k(s)),\,z_k'(s)\rangle ds.$$

Since $z'(s)=z_k(1)-z_k(0):=\Delta z_k$ and local smoothness condition in Lemma 16, we have

$$\begin{aligned}\mathcal{L}_\gamma(z_k(1))-\mathcal{L}_\gamma(z_k(0))&=\int_0^1\langle\nabla\mathcal{L}_\gamma(z_k(0)),\,\Delta z_k\rangle ds+\int_0^1\langle\nabla\mathcal{L}_\gamma(z_k(s))-\nabla\mathcal{L}_\gamma(z_k(0)),\,\Delta z_k\rangle ds\\&=\langle\nabla\mathcal{L}_\gamma(z_k(0)),\,\Delta z_k\rangle+\int_0^1\langle\nabla\mathcal{L}_\gamma(z_k(s))-\nabla\mathcal{L}_\gamma(z_k(0)),\,\Delta z_k\rangle ds\\&\le\langle\nabla\mathcal{L}_\gamma(z_k(0)),\,\Delta z_k\rangle+\int_0^1\|\nabla\mathcal{L}_\gamma(z_k(s))-\nabla\mathcal{L}_\gamma(z_k(0))\|_F\|\Delta z_k\|_F ds\\&\le\langle\nabla\mathcal{L}_\gamma(z_k(0)),\,\Delta z_k\rangle+L_k\int_0^1 s\|\Delta z_k\|_F^2 ds\\&\le\langle\nabla\mathcal{L}_\gamma(z_k(0)),\,\Delta z_k\rangle+\frac{L_k}{2}\|\Delta z_k\|_F^2\end{aligned}$$

Since $z_k(0)=z^k$ and $z_k(1)=z^{k+1}$, equivalently we have

$$\mathcal{L}_\gamma(\omega^{k+1},\theta^{k+1})\le\mathcal{L}_\gamma(\omega^k,\theta^k)+\langle\nabla\mathcal{L}_\gamma(\omega^k,\theta^k),\,z^{k+1}-z^k\rangle+\frac{L_k}{2}\|z^{k+1}-z^k\|_F^2.$$

where L_k is defined in (27). □

B.3.3 PROOF OF NON-UNIFORM POLYAK-ŁOJASIEWICZ (PL) CONDITION

Lemma 18 (Non-uniform PL condition of $g_y(u)$)**.** *Recall that $g_y(u) = -\log \mathbf{e}_y^\top \sigma(u)$ for any $u \in \mathbb{R}^V$, then*

$$\|\nabla g_y(u)\|_2^2 \geq 2\mu_g(u) g_y(u),$$

where

$$\mu_g(u) := \frac{V-1}{2V^2} \exp(-\sqrt{2}\|u\|_2).$$

Proof. Let $p = \sigma(u)$ and denote $p_y = [\sigma(u)]_y$. By direct calculation,

$$g_y(u) = -\log p_y, \qquad \nabla g_y(u) = p - \mathbf{e}_y.$$

Therefore,

$$\|\nabla g_y(u)\|_2^2 = (1-p_y)^2 + \sum_{v \neq y} p_v^2 \overset{(a)}{\geq} \frac{V}{V-1}(1-p_y)^2, \tag{28}$$

where in (a) we use $\sum_{v \neq y} p_v = 1 - p_y$ and the Cauchy–Schwarz inequality.

Next, for any $v \neq y$,

$$|u_v - u_y| = |(\mathbf{e}_v - \mathbf{e}_y)^\top u| \leq \sqrt{2}\|u\|_2 := a.$$

Therefore, we have

$$p_y = \frac{1}{1 + \sum_{v \neq y} \exp(u_v - u_y)} \in [p_{\min}, p_{\max}],$$

where

$$p_{\min} := \frac{1}{1 + (V-1)\exp(a)}, \qquad p_{\max} := \frac{1}{1 + (V-1)\exp(-a)}.$$

Define $f(p) := \frac{-\log p}{(1-p)^2}$. A direct calculation shows that f has one interior minimum on $(0,1)$. Hence its maximum over $[p_{\min}, p_{\max}]$ is attained at one of the endpoints:

$$\max_{p \in [p_{\min}, p_{\max}]} f(p) = \max\{f(p_{\min}), f(p_{\max})\}.$$

Moreover, the two endpoint values satisfy

$$f(p_{\min}) = \left(\frac{1}{V-1}\exp(-a) + 1\right)^2 \log\left(1 + (V-1)\exp(a)\right) \leq \frac{V^2}{(V-1)^2}(\log V + a),$$

$$f(p_{\max}) = \left(\frac{1}{V-1}\exp(a) + 1\right)^2 \log\left(1 + (V-1)\exp(-a)\right) \leq \frac{V^2}{V-1}\exp(a)$$

From direct circulation, for any $a \geq 0$ and $V \geq 2$, $\frac{V^2}{V-1}\exp(a) \geq \frac{V^2}{(V-1)^2}(\log V + a)$. Therefore, substituting $a = \sqrt{2}\|u\|_2$,

$$\frac{-\log p_y}{(1-p_y)^2} \leq \frac{V^2}{V-1}\exp(a) = \frac{V^2}{V-1}\exp(\sqrt{2}\|u\|_2).$$

Combining this with (28) gives

$$\|\nabla g_y(u)\|_2^2 \geq \frac{1}{V}\exp(-\sqrt{2}\|u\|_2) g_y(u).$$

This completes the proof. □

Lemma 19 (Non-uniform PL condition)**.** *Under Assumption 3, the offline penalized objective $\mathcal{L}_\gamma(\omega, \theta)$ satisfies a non-uniform PL condition. Specifically, if $\sigma_{\min}(H(\omega,\theta)) > 0$, then*

$$\|\nabla \mathcal{L}_\gamma(\omega,\theta)\|_F^2 \geq 2\mu_{\mathrm{PL}}(\omega,\theta)\mathcal{L}_\gamma(\omega,\theta),$$

where

$$\mu_{\mathrm{PL}}(\omega,\theta) := \sigma_{\min}^2(H(\omega,\theta))\, \underline{\mu}_g(\theta), \quad \underline{\mu}_g(\theta) := \frac{1}{2V}\exp(-\sqrt{2}B_X \sigma_{\max}(W_{\mathrm{ov}})). \tag{29}$$

Proof. By Lemma 18, for any single-token loss

$$g_y(u) := -\mathbf{e}_y^\top \log \sigma(u) = -\log[\sigma(u)]_y,$$

we have

$$\|\nabla g_y(u)\|_2^2 \geq 2\mu_g(u) g_y(u), \qquad \mu_g(u) = \frac{1}{2V}\exp(-\sqrt{2}\|u\|_2)$$

For each training token and validation token, by Assumption 3, we have

$$\|U_\theta(X_d^i)\|_2 \leq B_X \sigma_{\max}(W_{\mathrm{ov}}), \quad \|U_\theta(\widetilde{X}_d^j)\|_2 \leq B_X \sigma_{\max}(W_{\mathrm{ov}}).$$

Since $\mu_g(u)$ is non-increasing in $\|u\|_2$, we have

$$\mu_g(U_\theta(X_d^i)) \geq \underline{\mu}_g(\theta), \qquad \mu_g(U_\theta(\widetilde{X}_d^j)) \geq \underline{\mu}_g(\theta).$$

where $\mu_g(\theta) = \frac{1}{2V}\exp(-\sqrt{2}B_X\sigma_{\max}(W_{\mathrm{ov}}))$. Therefore,

$$\|\nabla g_{y_d^i}(U_\theta(X_d^i))\|_2^2 \geq 2\underline{\mu}_g(\theta) g_{y_d^i}(U_\theta(X_d^i)), \quad \|\nabla g_{\widetilde{y}_d^j}(U_\theta(\widetilde{X}_d^j))\|_2^2 \geq 2\underline{\mu}_g(\theta) g_{\widetilde{y}_d^j}(U_\theta(\widetilde{X}_d^j)). \tag{30}$$

Recall the weighted token-level feature matrix

$$H(\omega,\theta) := \left[\left\{\frac{1}{\sqrt{N'D}} h_\theta(\widetilde{X}_d^j)\right\}_{j\in[N'],\, d\in[D]}, \left\{\sqrt{\frac{\gamma\sigma_i(\omega)}{ND}} h_\theta(X_d^i)\right\}_{i\in[N],\, d\in[D]}\right],$$

and the weighted logit-gradient matrix

$$G(\omega,\theta) := \left[\left\{\frac{1}{\sqrt{N'D}} \nabla g_{\widetilde{y}_d^j}(U_\theta(\widetilde{X}_d^j))\right\}_{j\in[N'],\, d\in[D]}, \left\{\sqrt{\frac{\gamma\sigma_i(\omega)}{ND}} \nabla g_{y_d^i}(U_\theta(X_d^i))\right\}_{i\in[N],\, d\in[D]}\right].$$

Then

$$\begin{aligned}
\|G(\omega,\theta)\|_F^2 &= \frac{1}{N'D}\sum_{j=1}^{N'}\sum_{d=1}^{D}\|\nabla g_{\widetilde{y}_d^j}(U_\theta(\widetilde{X}_d^j))\|_2^2 + \frac{\gamma}{ND}\sum_{i=1}^{N}\sum_{d=1}^{D}\sigma_i(\omega)\|\nabla g_{y_d^i}(U_\theta(X_d^i))\|_2^2 \\
&\overset{(30)}{\geq} 2\underline{\mu}_g(\theta)\left[\frac{1}{N'D}\sum_{j=1}^{N'}\sum_{d=1}^{D} g_{\widetilde{y}_d^j}(U_\theta(\widetilde{X}_d^j)) + \frac{\gamma}{ND}\sum_{i=1}^{N}\sum_{d=1}^{D}\sigma_i(\omega) g_{y_d^i}(U_\theta(X_d^i))\right] \\
&= 2\underline{\mu}_g(\theta)\mathcal{L}_\gamma(\omega,\theta).
\end{aligned}$$

By the definitions of $G(\omega,\theta)$ and $H(\omega,\theta)$, the gradient with respect to W_{ov} admits the factorization

$$\nabla_{W_{\mathrm{ov}}}\mathcal{L}_\gamma(\omega,\theta) = G(\omega,\theta)H(\omega,\theta)^\top.$$

Therefore,

$$\begin{aligned}
\|\nabla_{W_{\mathrm{ov}}}\mathcal{L}_\gamma(\omega,\theta)\|_F^2 &= \|G(\omega,\theta)H(\omega,\theta)^\top\|_F^2 \\
&\geq \sigma_{\min}^2(H(\omega,\theta))\|G(\omega,\theta)\|_F^2 \\
&\geq 2\sigma_{\min}^2(H(\omega,\theta))\underline{\mu}_g(\theta)\mathcal{L}_\gamma(\omega,\theta).
\end{aligned}$$

Since W_{ov} is one block of θ, we have

$$\|\nabla\mathcal{L}_\gamma(\omega,\theta)\|_F^2 \geq \|\nabla_{W_{\mathrm{ov}}}\mathcal{L}_\gamma(\omega,\theta)\|_F^2 \geq 2\mu_{\mathrm{PL}}(\omega,\theta)\mathcal{L}_\gamma(\omega,\theta),$$

where $\mu_{\mathrm{PL}}(\omega,\theta)$ is defined in (29). □

B.3.4 Trajectory Control for the Local PL Region

For notational simplicity, along the offline PBGD trajectory, we write $H^k := H(\omega^k,\theta^k)$.

Lemma 20 (Trajectory and representation drift)**.** *Assume that $a \geq (N+N')D$ and H^0 has full column rank. Equivalently, there exists a constant $\sigma_0 > 0$ such that*

$$\sigma_{\min}(H^0) = \left(1+\frac{\gamma}{N}\right)^{1/2}\sigma_0 > 0. \tag{31}$$

Let $S_k := \sum_{t=0}^{k-1} \eta_t \mathcal{L}_\gamma(\omega^t, \theta^t)$, where $\eta_t > 0$ denotes the stepsize at iteration t. Along the offline PBGD trajectory,

$$\sigma_{\max}(W_{\mathrm{ov}}^k - W_{\mathrm{ov}}^0) \le \sqrt{2} B_X S_k, \quad \|\omega^k - \omega^0\|_2 \le \sqrt{2} S_k, \quad \sigma_{\max}(W_{\mathrm{kq}}^k - W_{\mathrm{kq}}^0) \le \frac{1}{2} B_X^3 B_{-1} S_k^2.$$

Moreover, suppose the initial representation matrix satisfies $\sigma_{\min}(H^0) = \left(1 + \frac{\gamma}{N}\right)^{1/2} \sigma_0$. Then

$$\sigma_{\min}(H^k) \ge \left(1 + \frac{\gamma}{N}\right)^{\frac{1}{2}} \left(\sigma_0 - \frac{1}{4} B_X^5 B_{-1}^2 S_k^2 - \left(\frac{\sqrt{2}}{4} B_X^4 B_{-1}^2 \sigma_{\max}(W_{\mathrm{ov}}^0) + \frac{1}{2} B_X \right) S_k \right)$$

Proof. **Step 1: Control of the W_{ov}, ω-trajectory.** Denote $S_k = \sum_{t=0}^{k-1} \eta_t \mathcal{L}_\gamma(\omega^t, \theta^t)$. By Lemma 15, for the trajectory of W_{ov}, we have

$$\sigma_{\max}\left(\nabla_{W_{\mathrm{ov}}} \mathcal{L}_\gamma(\omega, \theta)\right) \le \|\nabla_{W_{\mathrm{ov}}} \mathcal{L}_\gamma(\omega, \theta)\|_F \le \sqrt{2} B_X \mathcal{L}_\gamma(\omega, \theta)$$

We have one-step iteration inequality

$$\begin{aligned} \sigma_{\max}(W_{\mathrm{ov}}^{k+1} - W_{\mathrm{ov}}^k) &= \eta_k \sigma_{\max}(\nabla_{W_{\mathrm{ov}}} \mathcal{L}_\gamma(\omega^k, \theta^k)) \\ &\le \sqrt{2} B_X \eta_k \mathcal{L}_\gamma(\omega^k, \theta^k). \end{aligned}$$

The recursion telescopes after summing the one-step operator-norm bound, we have

$$\begin{aligned} \sigma_{\max}(W_{\mathrm{ov}}^k - W_{\mathrm{ov}}^0) &\le \sum_{t=0}^{k-1} \sigma_{\max}(W_{\mathrm{ov}}^{t+1} - W_{\mathrm{ov}}^t) \\ &\le \sqrt{2} B_X \sum_{t=0}^{k-1} \eta_t \mathcal{L}_\gamma(\omega^t, \theta^t) = \sqrt{2} B_X S_k. \end{aligned} \tag{32}$$

Using triangle inequality in (32), we have

$$\sigma_{\max}(W_{\mathrm{ov}}^k) \le \sqrt{2} B_X S_k + \sigma_{\max}(W_{\mathrm{ov}}^0). \tag{33}$$

For the trajectory of ω, since $\nabla_\omega \mathcal{L}_\gamma(\omega^k, \theta^k) = \frac{\gamma}{N} J_\sigma(\omega^k) \ell_{\mathrm{seq}}(\theta^k)$, denote $p^k := \sigma(\omega^k)$ and $q^k := (p^k)^\top \ell_{\mathrm{seq}}(\theta^k)$. Since

$$J_\sigma(\omega^k) \ell_{\mathrm{seq}}(\theta^k) = p^k \odot \ell_{\mathrm{seq}}(\theta^k) - p^k q^k,$$

the vector $J_\sigma(\omega^k) \ell_{\mathrm{seq}}(\theta^k)$ has zero sum. Moreover, its positive mass is bounded by q^k. Hence,

$$\|J_\sigma(\omega^k) \ell_{\mathrm{seq}}(\theta^k)\|_2 \le \sqrt{2} q^k.$$

Therefore,

$$\begin{aligned} \|\nabla_\omega \mathcal{L}_\gamma(\omega^k, \theta^k)\|_2 &\le \frac{\sqrt{2}\gamma}{N} \sum_{i=1}^N [\sigma(\omega^k)]_i \mathcal{L}_{\mathrm{SFT}}(\theta^k; X_i, y_i) \\ &\le \sqrt{2}\, \mathcal{L}_\gamma(\omega^k, \theta^k). \end{aligned}$$

The last inequality uses nonnegativity of the validation loss, so the weighted training loss is bounded by the full penalty objective.

Thus,

$$\begin{aligned} \|\omega^{k+1} - \omega^k\|_2 &= \eta_k \|\nabla_\omega \mathcal{L}_\gamma(\omega^k, \theta^k)\|_2 \\ &\le \sqrt{2}\, \eta_k \mathcal{L}_\gamma(\omega^k, \theta^k). \end{aligned}$$

Summing over $t = 0, \ldots, k-1$, we obtain

$$\|\omega^k - \omega^0\|_2 \le \sqrt{2} \sum_{t=0}^{k-1} \eta_t \mathcal{L}_\gamma(\omega^t, \theta^t) = \sqrt{2} S_k. \tag{34}$$

Step 2: Control of the W_{kq}-trajectory. By Lemma 15, for the trajectory of W_{kq}, we have

$$\sigma_{\max}(\nabla_{W_{\mathrm{kq}}} \mathcal{L}_\gamma(\omega, \theta)) \le \|\nabla_{W_{\mathrm{kq}}} \mathcal{L}_\gamma(\omega, \theta)\|_F \le \frac{\sqrt{2}}{2} B_X^2 B_{-1} \sigma_{\max}(W_{\mathrm{ov}}) \mathcal{L}_\gamma(\omega, \theta)$$

Then we have one-step iteration inequality

$$\sigma_{\max}(W_{\mathrm{kq}}^{k+1} - W_{\mathrm{kq}}^{k}) = \eta_k \sigma_{\max}(\nabla_{W_{\mathrm{kq}}} \mathcal{L}_\gamma(\omega^k, \theta^k)) \le \frac{\sqrt{2}}{2} B_X^2 B_{-1} \sigma_{\max}(W_{\mathrm{ov}}^k) \mathcal{L}_\gamma(\omega^k, \theta^k).$$

The recursion telescopes after summing the one-step operator-norm bound and use 33, we have

$$\begin{aligned}
\sigma_{\max}(W_{\mathrm{kq}}^{k} - W_{\mathrm{kq}}^{0}) &\le \frac{\sqrt{2}}{2} B_X^2 B_{-1} \sum_{t=0}^{k-1} \sigma_{\max}(W_{\mathrm{ov}}^t) \eta_t \mathcal{L}_\gamma(\omega^t, \theta^t) \\
&\le \frac{\sqrt{2}}{2} B_X^2 B_{-1} \sum_{t=0}^{k-1} \left(\sqrt{2} B_X S_t + \sigma_{\max}(W_{\mathrm{ov}}^0) \right) \eta_t \mathcal{L}_\gamma(\omega^t, \theta^t) \\
&\overset{(a)}{=} \frac{\sqrt{2}}{2} B_X^2 B_{-1} \left(\frac{\sqrt{2}}{2} B_X S_k^2 + \sigma_{\max}(W_{\mathrm{ov}}^0) S_k - \frac{\sqrt{2}}{2} B_X \sum_{t=0}^{k-1} \eta_t^2 \mathcal{L}_\gamma(\omega^t, \theta^t)^2 \right) \\
&\le \frac{1}{2} B_X^2 B_{-1} \left(B_X S_k^2 + \sqrt{2} \sigma_{\max}(W_{\mathrm{ov}}^0) S_k \right) \qquad (35)
\end{aligned}$$

where (a) is the identity $2\sum_{s<t} a_s a_t = (\sum_t a_t)^2 - \sum_t a_t^2$, applied with $a_t = \eta_t \mathcal{L}_\gamma(\omega^t, \theta^t)$.

Step 3: Control of the H-trajectory.

Recall that

$$H(\omega, \theta) := \left[\left\{ \frac{1}{\sqrt{N'D}} h_\theta(\widetilde{X}_d^j) \right\}_{j \in [N'],\, d \in [D]}, \left\{ \sqrt{\frac{\gamma \sigma_i(\omega)}{ND}}\, h_\theta(X_d^i) \right\}_{i \in [N],\, d \in [D]} \right] \in \mathbb{R}^{a \times (N+N')D}.$$

Then

$$\|H(\widetilde{\omega}, \widetilde{\theta}) - H(\omega, \theta)\|_F \le \|H(\widetilde{\omega}, \widetilde{\theta}) - H(\omega, \widetilde{\theta})\|_F + \|H(\omega, \widetilde{\theta}) - H(\omega, \theta)\|_F.$$

For the θ-dependent term, by Lemma 10, we have

$$\begin{aligned}
\|h_{\widetilde{\theta}}(\widetilde{X}_d^j) - h_\theta(\widetilde{X}_d^j)\|_2 &\le \frac{1}{2} B_X^2 B_{-1} \sigma_{\max}(\widetilde{W}_{\mathrm{kq}} - W_{\mathrm{kq}}), \\
\|h_{\widetilde{\theta}}(X_d^i) - h_\theta(X_d^i)\|_2 &\le \frac{1}{2} B_X^2 B_{-1} \sigma_{\max}(\widetilde{W}_{\mathrm{kq}} - W_{\mathrm{kq}}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|H(\omega, \widetilde{\theta}) - H(\omega, \theta)\|_F^2 &= \frac{1}{N'D} \sum_{j=1}^{N'} \sum_{d=1}^{D} \|h_{\widetilde{\theta}}(\widetilde{X}_d^j) - h_\theta(\widetilde{X}_d^j)\|_2^2 \\
&\quad + \frac{\gamma}{ND} \sum_{i=1}^{N} \sum_{d=1}^{D} \sigma_i(\omega) \|h_{\widetilde{\theta}}(X_d^i) - h_\theta(X_d^i)\|_2^2 \\
&\le \frac{1}{4} B_X^4 B_{-1}^2 \left(1 + \frac{\gamma}{N} \right) \sigma_{\max}(\widetilde{W}_{\mathrm{kq}} - W_{\mathrm{kq}})^2.
\end{aligned}$$

Equivalently, we have

$$\|H(\omega, \widetilde{\theta}) - H(\omega, \theta)\|_F \le \frac{1}{2} B_X^2 B_{-1} \left(1 + \frac{\gamma}{N} \right)^{1/2} \sigma_{\max}(\widetilde{W}_{\mathrm{kq}} - W_{\mathrm{kq}}). \qquad (36)$$

For the ω-dependent term, the validation columns do not depend on ω. Thus

$$\begin{aligned}
\|H(\widetilde{\omega}, \widetilde{\theta}) - H(\omega, \widetilde{\theta})\|_F^2 &= \frac{\gamma}{ND} \sum_{i=1}^{N} \sum_{d=1}^{D} \left(\sqrt{\sigma_i(\widetilde{\omega})} - \sqrt{\sigma_i(\omega)} \right)^2 \|h_{\widetilde{\theta}}(X_d^i)\|_2^2 \\
&\le \frac{\gamma}{N} B_X^2 \left\| \sqrt{\sigma(\widetilde{\omega})} - \sqrt{\sigma(\omega)} \right\|_2^2.
\end{aligned}$$

Since the map $r(\omega) := \sqrt{\sigma(\omega)}$ is globally $\sqrt{2}/4$-Lipschitz, we obtain

$$\left\| \sqrt{\sigma(\widetilde{\omega})} - \sqrt{\sigma(\omega)} \right\|_2 \le \frac{\sqrt{2}}{4} \|\widetilde{\omega} - \omega\|_2.$$

Therefore,

$$\|H(\widetilde{\omega},\widetilde{\theta}) - H(\omega,\widetilde{\theta})\|_F \le \frac{\sqrt{2}}{4}\sqrt{\frac{\gamma}{N}} B_X \|\widetilde{\omega} - \omega\|_2. \tag{37}$$

Combining the two estimates (36) and (37) gives

$$\begin{aligned}
&\sigma_{\max}(H(\widetilde{\omega},\widetilde{\theta}) - H(\omega,\theta)) \\
&\le \|H(\widetilde{\omega},\widetilde{\theta}) - H(\omega,\theta)\|_F \\
&\le \frac{1}{2}B_X^2 B_{-1}\left(1+\frac{\gamma}{N}\right)^{1/2}\sigma_{\max}(\widetilde{W}_{\mathrm{kq}} - W_{\mathrm{kq}}) + \frac{\sqrt{2}}{4}\sqrt{\frac{\gamma}{N}} B_X\|\widetilde{\omega}-\omega\|_2 \\
&\le \left(1+\frac{\gamma}{N}\right)^{1/2}\left(\frac{1}{2}B_X^2 B_{-1}\sigma_{\max}(\widetilde{W}_{\mathrm{kq}} - W_{\mathrm{kq}}) + \frac{\sqrt{2}}{4} B_X\|\widetilde{\omega}-\omega\|_2\right).
\end{aligned}$$

Taking $(\widetilde{\omega},\widetilde{\theta}) = (\omega^k,\theta^k)$ and $(\omega,\theta) = (\omega^0,\theta^0)$, and using the trajectory bounds,

$$\sigma_{\max}(W_{\mathrm{kq}}^k - W_{\mathrm{kq}}^0) \le \frac{1}{2}B_X^3 B_{-1} S_k^2 + \frac{\sqrt{2}}{2}B_X^2 B_{-1}\sigma_{\max}(W_{\mathrm{ov}}^0) S_k, \quad \|\omega^k - \omega^0\|_2 \le \sqrt{2} S_k,$$

we obtain

$$\begin{aligned}
\sigma_{\max}(H^k - H^0) &\le \left(1+\frac{\gamma}{N}\right)^{1/2}\left(\frac{1}{2}B_X^2 B_{-1}\sigma_{\max}(W_{\mathrm{kq}}^k - W_{\mathrm{kq}}^0) + \frac{\sqrt{2}}{4}B_X\|\omega^k-\omega^0\|_2\right) \\
&\le \left(1+\frac{\gamma}{N}\right)^{1/2}\left(\frac{1}{4}B_X^5 B_{-1}^2 S_k^2 + \left(\frac{\sqrt{2}}{4}B_X^4 B_{-1}^2\sigma_{\max}(W_{\mathrm{ov}}^0) + \frac{1}{2}B_X\right) S_k\right).
\end{aligned}$$

Under condition (31), we have

$$\begin{aligned}
\sigma_{\min}(H^k) &\ge \sigma_{\min}(H^0) - \sigma_{\max}(H^k - H^0) \\
&\ge \left(1+\frac{\gamma}{N}\right)^{\frac{1}{2}}\left(\sigma_0 - \frac{1}{4}B_X^5 B_{-1}^2 S_k^2 - \left(\frac{\sqrt{2}}{4}B_X^4 B_{-1}^2\sigma_{\max}(W_{\mathrm{ov}}^0) + \frac{1}{2}B_X\right) S_k\right)
\end{aligned}$$

□

B.3.5 Proof of Theorem 2

Lemma 21 (Descent recursion under shifted local PL and smoothness)**.** *Under Assumption 3, let*

$$\begin{aligned}
\mu_k &= \frac{\sigma_{\min}(H^k)^2}{2V}\exp(-\sqrt{2}B_X\sigma_{\max}(W_{\mathrm{ov}}^k)) \\
L_k &= 4B_X^2(1+B_{-1})\left(1+B_X B_{-1}\sigma_{\max}(W_{\mathrm{ov}}^k)\right)^2\left(1+\frac{\gamma}{N}\right) \\
&\quad + \frac{3\gamma}{\sqrt{N}}\left(\log V + B_X + B_X\left(1+B_X B_{-1}\right)\sigma_{\max}(W_{\mathrm{ov}}^k)\right)
\end{aligned}$$

Suppose the step size satisfies $\eta_k \le 1/L_k$ and the shifted local PL bound

$$\mathcal{L}_\gamma(\omega^{k+1},\theta^{k+1}) \le (1-\eta_k\mu_k)\mathcal{L}_\gamma(\omega^k,\theta^k),$$

Proof. It follows directly from Lemma 17 and Lemma 19. □

Theorem 22 (Complete version of Theorem 2)**.** *Consider the offline PBGD updates in (5) for the one-layer softmax Transformer in (4). Suppose Assumption 1 holds, and additionally, assume the upper-level dataset is separable. Assume Assumption 3 holds and $a \ge (N+N')D$. Let $k_0 \ge 0$ be a chosen burn-in index and fix a target accuracy $\epsilon > 0$. Set the penalty parameter as $\gamma = \epsilon^{-1/2}$ and define*

$$R_{k_0} := \sigma_{\max}(W_{\mathrm{ov}}^{k_0}), \quad \sigma_{\min}(H^{k_0}) = \left(1+\frac{\gamma}{N}\right)^{1/2}\sigma_{k_0}, \quad \mathcal{L}_\gamma(\omega^{k_0},\theta^{k_0}) = \left(1+\frac{\gamma}{N}\right) l^{k_0}$$

and assume the burn-in representation is non-degenerate, i.e., $\sigma_{k_0} > 0$.

Define the certified representation radius

$$S_H(k_0) := \frac{-\left(\frac{\sqrt{2}}{2}B_X^4 B_{-1}^2 R_{k_0} + B_X\right) + \sqrt{\left(\frac{\sqrt{2}}{2}B_X^4 B_{-1}^2 R_{k_0} + B_X\right)^2 + 2B_X^5 B_{-1}^2 \sigma_{k_0}}}{B_X^5 B_{-1}^2}.$$

Assume that at the burn-in step k_0, we have

$$\frac{8Vl^{k_0}\exp\left(\sqrt{2}B_X R_{k_0}\right)}{\sigma_{k_0}^2} \le M_H(k_0), \tag{38}$$

where $M_H(k_0) := \max_{0<S\le S_H(k_0)} S\exp(-2B_X^2 S)$. Then the set

$$\mathcal{S}(k_0) = \{S \in (0, S_H(k_0)] : \frac{8Vl^{k_0}}{\sigma_{k_0}^2\exp(-2B_X^2 S - \sqrt{2}B_X R_{k_0})} \le S\} \tag{39}$$

is non-empty. We define $S_\epsilon(k_0) = \sup_{S\in\mathcal{S}(k_0)} S$. And further define

$$\mu_\epsilon(k_0) := \left(1+\frac{\gamma}{N}\right)\frac{\sigma_{k_0}^2}{8V}\exp\left(-2B_X^2 S_\epsilon(k_0) - \sqrt{2}B_X R_{k_0}\right), \tag{40}$$

$$\begin{aligned} L_\epsilon(k_0) :=\ & 4B_X^2(1+B_{-1})\left(1 + B_X B_{-1} R_{k_0} + \sqrt{2}B_X^2 B_{-1} S_\epsilon(k_0)\right)^2\left(1+\frac{\gamma}{N}\right) \\ & + \frac{3\gamma}{\sqrt{N}}\left(\log V + B_X + B_X(1+B_X B_{-1})R_{k_0} + \sqrt{2}B_X^2(1+B_X B_{-1})S_\epsilon(k_0)\right). \end{aligned} \tag{41}$$

For every $t \ge k_0$, choose the fixed stepsize $\eta_t = \eta := \frac{1}{L_\epsilon(k_0)}$.

Define

$$A(k_0) := \log\left(1 + (V-1)\exp\left(-2B_X^2 S_\epsilon(k_0) - \sqrt{2}B_X R_{k_0}\right)\right), \tag{42}$$

For any $\epsilon > \epsilon_0$, where

$$\epsilon_0 = \max\left\{2A(k_0), \left(\frac{2A(k_0)}{N}\right)^{2/3}\right\}, \tag{43}$$

the iterates of the offline BDS *satisfy $\tau_\epsilon(k_0) := \inf\left\{k \ge k_0 : \mathcal{L}_\gamma(\omega^k, \theta^k) \le \epsilon\right\}$, and*

$$\tau_\epsilon(k_0) - k_0 \le \left\lceil \frac{1}{\eta\mu_\epsilon(k_0)} \log\frac{\mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0})}{\epsilon}\right\rceil < \infty. \tag{44}$$

Proof. We prove the result by showing that the trajectory remains inside the certified local region determined by $S_\epsilon(k_0)$, where both the local PL constant and the smoothness constant are uniformly controlled.

For $k \ge k_0$, define the accumulated trajectory length

$$S_{k,k_0} := \sum_{t=k_0}^{k-1} \eta\,\mathcal{L}_\gamma(\omega^t, \theta^t).$$

Suppose first that $S_{k,k_0} \le S_\epsilon(k_0)$. By Lemma 20, applied from the burn-in point k_0, we have

$$\sigma_{\max}(W_{\text{ov}}^k) \le R_{k_0} + \sqrt{2}B_X S_{k,k_0}, \tag{45}$$

$$\sigma_{\min}(H^k) \ge \left(1+\frac{\gamma}{N}\right)^{1/2}\left[\sigma_{k_0} - \frac{1}{4}B_X^5 B_{-1}^2 S_{k,k_0}^2 - \left(\frac{\sqrt{2}}{4}B_X^4 B_{-1}^2 R_{k_0} + \frac{1}{2}B_X\right)S_{k,k_0}\right] \tag{46}$$

Since $S_{k,k_0} \le S_\epsilon(k_0) \le S_H(k_0)$, the definition of $S_H(k_0)$ gives

$$\frac{1}{4}B_X^5 B_{-1}^2 S_{k,k_0}^2 + \left(\frac{\sqrt{2}}{4}B_X^4 B_{-1}^2 R_{k_0} + \frac{1}{2}B_X\right)S_{k,k_0} \le \frac{1}{2}\sigma_{k_0}.$$

Therefore,

$$\sigma_{\min}(H^k) \ge \frac{1}{2}\left(1+\frac{\gamma}{N}\right)^{1/2}\sigma_{k_0}. \tag{47}$$

Similarly, (45) gives

$$\sigma_{\max}(W_{\text{ov}}^k) \le R_{k_0} + \sqrt{2} B_X S_\epsilon(k_0). \tag{48}$$

Combining (47) and (48) with Lemma 19, we obtain the uniform shifted local PL bound inside the certified region:

$$\|\nabla \mathcal{L}_\gamma(\omega^k, \theta^k)\|_F^2 \ge 2\mu_\epsilon(k_0) \mathcal{L}_\gamma(\omega^k, \theta^k), \tag{49}$$

where $\mu_\epsilon(k_0)$ is defined in (40).

Moreover, by Lemma 17 and (48), the local smoothness constant satisfies $L_k \le L_\epsilon(k_0)$. Since $\eta = 1/L_\epsilon(k_0)$, we have $\eta \le 1/L_k$ in the certified region. Hence Lemma 21 gives

$$\mathcal{L}_\gamma(\omega^{k+1}, \theta^{k+1}) \le (1 - \eta\mu_\epsilon(k_0))\, \mathcal{L}_\gamma(\omega^k, \theta^k). \tag{50}$$

On the other hand, iterating (50) yields

$$\mathcal{L}_\gamma(\omega^k, \theta^k) \le (1 - \eta\mu_\epsilon(k_0))^{k-k_0}\, \mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0}) \tag{51}$$

Consequently,

$$\begin{aligned} S_{k,k_0} &= \eta \sum_{t=k_0}^{k-1} \mathcal{L}_\gamma(\omega^t, \theta^t) \\ &\le \eta \sum_{t=k_0}^{k-1} (1 - \eta\mu_\epsilon(k_0))^{t-k_0}\, \mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0}) \\ &\le \frac{\mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0})}{\mu_\epsilon(k_0)} \overset{(a)}{\le} S_\epsilon(k_0), \end{aligned}$$

where (a) is from the burn-in initialization condition (38); indeed, by the definitions of $\mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0})$ and $\mu_\epsilon(k_0)$,

$$\frac{\mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0})}{\mu_\epsilon(k_0)} = \frac{8Vl^{k_0} \exp\left(2B_X^2 S_\epsilon(k_0) + \sqrt{2} B_X R_{k_0}\right)}{\sigma_{k_0}^2} \le S_\epsilon(k_0).$$

Therefore, by induction, the whole trajectory remains inside the certified tube

$$S_{k,k_0} \le S_\epsilon(k_0), \qquad k \ge k_0.$$

For any step $t \ge k_0$ such that $\sigma_{\max}(W_{\text{ov}}^t) \le \sigma_{\max}(W_{\text{ov}}^{k_0}) + \sqrt{2} B_X S_\epsilon(k_0)$, the lower bound $\mathcal{L}_\gamma(\omega^t, \theta^t) \ge \left(1 + \frac{\gamma}{N}\right) A(k_0) := \mathcal{L}_\gamma^*(k_0)$ where $A(k_0)$ is defined in (42).

With the penalty choice $\gamma = \epsilon^{-1/2}$, by the condition (43),

$$A(k_0) \le \frac{\epsilon}{2}, \qquad \frac{A(k_0)}{N\sqrt{\epsilon}} \le \frac{\epsilon}{2}.$$

Consequently,

$$\mathcal{L}_\gamma^*(k_0) = \left(1 + \frac{\gamma}{N}\right) A(k_0) = A(k_0) + \frac{A(k_0)}{N\sqrt{\epsilon}} \le \epsilon.$$

When the condition (43) holds true, the uniform PL (40) and smoothness (41) bounds above hold.

Now, from (51) and the inequality $1 - x \le e^{-x}$, we have

$$\mathcal{L}_\gamma(\omega^k, \theta^k) \le \exp\left(-\eta\mu_\epsilon(k_0)(k - k_0)\right) \mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0}).$$

Thus it is sufficient to take

$$k - k_0 \ge \left\lceil \frac{1}{\eta\mu_\epsilon(k_0)} \log \frac{\mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0})}{\epsilon} \right\rceil.$$

This proves the hitting-time bound

$$\tau_\epsilon(k_0) - k_0 \le \left\lceil \frac{1}{\eta\mu_\epsilon(k_0)} \log \frac{\mathcal{L}_\gamma(\omega^{k_0}, \theta^{k_0})}{\epsilon} \right\rceil < \infty.$$

□

B.4 Proof of Theorem 3

Before proceeding, we first define the bilevel linear scalarization problem as

$$\texttt{BLS}: \min_{\lambda\in\Delta^N,\theta} \frac{1}{N'}\sum_{i=1}^{N'} \mathcal{L}_{\text{val}}(\theta;\tilde{x}^i,\tilde{y}^i), \ \text{ s.t. } \theta\in\mathcal{S}(\lambda) := \arg\min_{\theta'} \frac{1}{N}\sum_{i=1}^{N}\lambda_i\mathcal{L}_{\text{SFT}}(\theta';x^i,y^i) \quad (52)$$

where $\Delta^N := \{\lambda\in\mathbb{R}^N : \lambda_i\geq 0, \sum_{i=1}^N\lambda_i = 1\}$ is the simplex. Clearly, by setting $\lambda_i = \sigma_i(\omega)$, BLS is equivalent to BDS. So the remaining task is to prove the equivalence of BLS and BMO.

Proof. Define $f_i(\theta) := \mathcal{L}_{\text{SFT}}(\theta;x^i,y^i)$ for each $i\in[N]$, and write $f(\theta) := (f_1(\theta),\dots,f_N(\theta))$. By switching the optimization order in the upper level, the equivalence of BLS and BMO reduces to proving

$$\text{WP}(f) = \bigcup_{\lambda\in\Delta^N}\mathcal{S}(\lambda).$$

By Assumption 1, there exists $\bar{\theta}\in\mathbb{R}^h$ such that

$$f_i(\bar{\theta}) = 0, \qquad \forall i\in[N].$$

Since each $f_i(\theta)\geq 0$, for any $\lambda\in\Delta^N$ the scalarized lower-level objective $\frac{1}{N}\sum_{i=1}^N\lambda_i f_i(\theta)$ is nonnegative and reaches value 0 at $\bar{\theta}$. Therefore,

$$\begin{aligned}\theta\in\mathcal{S}(\lambda) &\iff \frac{1}{N}\sum_{i=1}^N\lambda_i f_i(\theta) = 0\\ &\iff f_i(\theta) = 0 \quad \text{for every } i \text{ with } \lambda_i > 0. \end{aligned} \quad (53)$$

Hence

$$\bigcup_{\lambda\in\Delta^N}\mathcal{S}(\lambda) = \left\{\theta\in\mathbb{R}^h : \exists i\in[N] \text{ such that } f_i(\theta) = 0\right\}. \quad (54)$$

We next characterize the weak Pareto set of f. First, suppose $f_i(\theta) > 0$ for every $i\in[N]$. Then $\bar{\theta}$ strictly improves all coordinates because

$$f_i(\bar{\theta}) = 0 < f_i(\theta), \qquad \forall i\in[N].$$

Hence $\theta\notin\text{WP}(f)$.

Conversely, if $f_j(\theta) = 0$ for some $j\in[N]$, then there does not exist any θ' such that

$$f_j(\theta') < f_j(\theta) = 0,$$

because every SFT loss is nonnegative. Therefore no θ' can strictly improve all coordinates simultaneously, which implies $\theta\in\text{WP}(f)$. We have thus proved

$$\text{WP}(f) = \left\{\theta\in\mathbb{R}^h : \exists i\in[N] \text{ such that } f_i(\theta) = 0\right\}. \quad (55)$$

Combining (54) and (55), we obtain

$$\text{WP}(f) = \bigcup_{\lambda\in\Delta^N}\mathcal{S}(\lambda).$$

Therefore BMO and BLS minimize the same validation objective over the same feasible set, so their global minimizers coincide. Since BLS is equivalent to BDS through the change of variables $\lambda_i = \sigma_i(\omega)$, the global solutions of BMO and BDS coincide. □

C Implicit weight assigned by BMO

In this section, we provide the proof of Lemma 4. First we derive the generalized implicit weight assigned by bilevel multi-objective optimization problem and then specify each objective as the response-level SFT loss so that we can obtain Lemma 4.

In previous section, we showed that the optimal LLM model θ given by BDS and BMO are the same. This inspires the question that

How to interpret the implicit weights assigned by algorithm for solving BMO?

To solve BMO in (6), we consider converting weak Pareto set optimization into a scalar objective. In (Tanabe et al., 2024), it is shown that weakly Pareto set $\mathrm{WP}(\theta)$ for any lower semicontinuous multi-function $\mathcal{L}(\theta)$ can be equivalently expressed by a merit function

$$u(\theta) = \sup_{\theta'} \min_{m\in[M]} \{\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta')\} \tag{56}$$

where $u(\theta) \geq 0$ and the equality holds if and only if $\theta \in \mathrm{WP}(\theta)$. This merit function is easy to check by the definition of weak Pareto set in Definition 1 as

$$\begin{aligned}\mathrm{WP}(\mathcal{L}) &= \{\theta \mid \nexists \theta' \text{ s.t. } \forall m \in [M], \mathcal{L}_m(\theta') < \mathcal{L}_m(\theta)\} \\ &= \{\theta \mid \forall \theta' \text{ s.t. } \exists m \in [M], \mathcal{L}_m(\theta') \geq \mathcal{L}_m(\theta)\}\end{aligned}$$

while the latter can be expressed in terms of the supremum and infimum in (56). Consequently, BMO can be reformulated to $\min_\theta \mathcal{L}_{\mathrm{val}}(\theta),\ \text{s.t. } u(\theta) \leq 0$, and can be solved sequentially by its penalty reformulation (Chen et al., 2025)

$$\texttt{PMO}: \min_{\theta}\ \mathcal{L}_{\mathrm{val}}(\theta) + \gamma_k u(\theta) \tag{57}$$

with enlarging $\gamma_k \to \infty$ similar to BDS. Then the bottleneck of designing a gradient-based algorithm on (57) is $u(x)$, which is usually non-differentiable due to $\max$ operator.

One popular choice to estimate $u(x)$ is to rewrite it in a $\max$ form and estimate the $\max$ operator by log-sum-exponential (LSE) function (Rockafellar & Wets, 2009)

$$\mathrm{LSE}(q;\tau) = \frac{1}{\tau}\log\left(\sum_{m=1}^{M}\exp\left(\tau q_m\right)\right).$$

It is well-known that for any vector $q \in \mathbb{R}^M$, $\max_{m\in[M]} q_m \leq \mathrm{LSE}(q;\tau) \leq \max_{m\in[M]} q_m + \frac{\log M}{\tau}$ (Boyd & Vandenberghe, 2004; Rockafellar & Wets, 2009). Therefore, $\min_{m\in[M]} q_m \approx -\mathrm{LSE}(-q;\tau)$ can be approximated by a smooth function. Applying these results into PMO in (57) leads to

$$\texttt{PMO}_{\mathrm{LSE}}: \min_{\theta}\ \mathcal{L}_{\mathrm{val}}(\theta) - \frac{\gamma_k}{\tau}\inf_{\theta'}\mathrm{LSE}\left(\mathcal{L}(\theta') - \mathcal{L}(\theta);\tau\right) \tag{58}$$

When leveraging $\texttt{PMO}_{\mathrm{LSE}}$ to solve the BDS problem, the formulation can be further simplified.

Lemma 23. *Letting $\mathcal{L}_m(\theta) = \mathcal{L}_{SFT}(\theta; x^m, y^m)$ and under Assumption 1, (58) can be simplified as*

$$\mathit{PMO}^{\mathrm{S}}_{\mathrm{LSE}}: \min_{\theta}\ \mathcal{L}_{\mathrm{val}}(\theta) - \frac{\gamma_k}{\tau}\mathrm{LSE}(-\mathcal{L}(\theta);\tau) \tag{59}$$

where $\mathcal{L}(\theta) := \left[\mathcal{L}_{SFT}(\theta; x^1, y^1), \cdots, \mathcal{L}_{SFT}(\theta; x^M, y^M)\right]$ is the per-sample SFT loss vector.

Proof. Letting θ^* be the shared minimizer in Assumption 1, we first want to prove

$$\sup_{\theta'}\min_{m\in[M]}\left(\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta')\right) = \min_{m\in[M]}\mathcal{L}_m(\theta). \tag{60}$$

which is equivalent to

$$\sup_{\theta'}\min_{\lambda\in\Delta^M}\sum_{m=1}^{M}\lambda_m\left(\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta')\right) = \min_{\lambda\in\Delta^M}\sum_{m=1}^{M}\lambda_m\mathcal{L}_m(\theta)$$

because $\min_{m\in[M]} q_m = \min_{\lambda\in\Delta^M}\lambda_m q_m$.

First we have

$$\sup_{\theta'}\min_{\lambda\in\Delta^M}\sum_{m=1}^{M}\lambda_m\left(\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta')\right) \geq \min_{\lambda\in\Delta^M}\sum_{m=1}^{M}\lambda_m\left(\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta^*)\right) = \min_{\lambda\in\Delta^M}\sum_{m=1}^{M}\lambda_m\mathcal{L}_m(\theta). \tag{61}$$

On the other hand, for any $m \in [M]$,

$$\sup_{\theta'}\min_{\lambda\in\Delta^M}\sum_{m=1}^{M}\lambda_m\left(\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta')\right) \leq \sup_{\theta'}\left(\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta')\right) = \mathcal{L}_m(\theta) \tag{62}$$

Since (62) holds for any $m \in [M]$,

$$\sup_{\theta'}\min_{\lambda\in\Delta^M}\sum_{m=1}^{M}\lambda_m\left(\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta')\right) \leq \min_{m\in[M]}\mathcal{L}_m(\theta) = \min_{\lambda\in\Delta^M}\sum_{m=1}^{M}\lambda_m\mathcal{L}_m(\theta). \tag{63}$$

where the last equality is because $\min_{m\in[M]} q_m = \min_{\lambda\in\Delta^M} \sum_{m=1}^M q_m$. Then with (61), we know

$$\sup_{\theta'} \min_{\lambda\in\Delta^M} \sum_{m=1}^M \lambda_m \left(\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta')\right) = \min_{\lambda\in\Delta^M} \sum_{m=1}^M \lambda_m \mathcal{L}_m(\theta).$$

so that (60) holds. Then

$$\inf_{\theta'} \mathrm{LSE}\left(\mathcal{L}(\theta') - \mathcal{L}(\theta); \tau\right) \le \mathrm{LSE}\left(\mathcal{L}(\theta^*) - \mathcal{L}(\theta); \tau\right) = \mathrm{LSE}\left(-\mathcal{L}(\theta); \tau\right). \tag{64}$$

On the other hand,

$$\begin{aligned}\inf_{\theta'} \mathrm{LSE}\left(\mathcal{L}(\theta') - \mathcal{L}(\theta); \tau\right) &\overset{(a)}{\ge} \inf_{\theta'} \left\{ \max_{m\in[M]} \left(\mathcal{L}_m(\theta') - \mathcal{L}_m(\theta)\right) \right\} \overset{(b)}{=} \max_{m\in[M]} -\mathcal{L}_m(\theta) \\ &\ge \mathrm{LSE}\left(-\mathcal{L}(\theta); \tau\right) - \frac{\log M}{\tau} \end{aligned} \tag{65}$$

where (a) is because $\max_{m\in[M]} q_m \le \mathrm{LSE}(q;\tau)$, (b) is earned by (60) and $\min_{m\in[M]} q_m = \min_{\lambda\in\Delta^M} \sum_{m=1}^M q_m$, and (c) is because $\mathrm{LSE}(q;\tau) \le \max_{m\in[M]} q_m + \frac{\log M}{\tau}$. Then since $\frac{\log M}{\tau}$ is a constant, we get the conclusion. □

Implicit BMO weights and their relation to importance ratios when $\tau = 1$. Taking the gradient over $\texttt{PMO}^{\mathrm{S}}_{\mathrm{LSE}}$ in (59) gives the update direction as

$$\nabla \mathcal{L}_{\mathrm{val}}(\theta) + \gamma_k \sum_{m=1}^M \lambda_m \nabla \mathcal{L}_{\mathrm{SFT}}(\theta; x^m, y^m) \tag{66}$$

where $\lambda_m = \frac{\exp(-\tau \mathcal{L}_{\mathrm{SFT}}(\theta; x^m, y^m))}{\sum_{i=1}^M \exp(-\tau \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^i))}$ is given by the softmax policy of negative per-sample SFT loss. Compared with (10a), instead of directly learning the data weight, $\texttt{PMO}^{\mathrm{S}}_{\mathrm{LSE}}$ chooses a special logit of data weight as

$$\omega^*(\theta) = -\tau \left[\mathcal{L}_{\mathrm{SFT}}(\theta; x^1, y^1), \cdots, \mathcal{L}_{\mathrm{SFT}}(\theta; x^M, y^M)\right] \tag{67}$$

which is also inversely determined by the per-sample SFT loss. If the per-sample SFT loss of a data sample remains low after joint descent with the validation function, it is likely drawn from the same distribution as the validation set, and thus we assign a higher weight to this sample.

Application to Lemma 4. Consider i-th question and let $y^m, m \in [G]$ be the group of generations to i-th question. Then the implicit weight of g-th generation is

$$\lambda_{i,g} = \mathrm{softmax}(-\tau \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^{i,g}_{\mathrm{old}})) \propto \exp(-\tau \mathcal{L}_{\mathrm{SFT}}(\theta; x^i, y^{i,g}_{\mathrm{old}})) = \pi_\theta(x^i, y^{i,g}_{\mathrm{old}})$$

which is proportional to the importance ratio of g-th response when the old-policy likelihoods are approximately equal. In this way, we are not only assigning question-level validation score, but also enforce response-level validation score to prioritize response adhere to the validation dataset in the proposed algorithm. In experiments, we test the cosine similarity of important-ratio weights and the BMO weights and it achieves 0.984 similarity, which further validates this practical implementation choice.

D Additional Experiments

D.1 Details of Experimental Setup

In this section, we present the detailed experimental setup and the hyperparameter choices.

General Setup. For both fine-tuning tasks, we use 56000 samples for the LLAMA-3-8B (Grattafiori et al., 2024) model and 9600 samples for the PYTHIA-1b (Biderman et al., 2023) model. We only process the question with length shorter than 2048. Both models are adapted with Low-Rank Adaptation (LoRA) (ALPHA 16, RANK 16). The learning rates are set to 5×10^{-6} for the LoRA parameter θ update and 1×10^{-4} for the selector ω update, using Adam (Kingma & Ba, 2015) as the optimizer. Fine-tuning was performed using PyTorch with the DeepSpeed library `https://github.com/deepspeedai/DeepSpeed` to optimize memory usage. We use effective train batch size 32 and micro batch size 8 for both tasks, and use zero stage 2 with gradient checkpointing in DeepSpeed. We train 3 epochs for all algorithms.

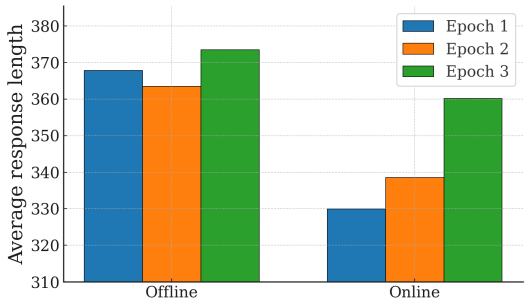

Figure 4: Average response length of the top-10% questions by learned weight from offline BDS and online selection. Online selection shifts from shorter- to longer-response questions over training.

Algorithm hyperparameters. We use penalty constant of $\gamma_k = \frac{\rho_k}{1-\rho_k}$ with ρ_k initialized as 0.1 and increased by 0.1 after every epoch for both offline and online selection, as suggested by (Shen et al., 2025a). For the online selection approach, we generate responses to the masked questions using a batch size of 64 for the first quality enhancement task and 16 for the second safety-aware fine-tuning task. We generate new responses every $K_{\text{gen}} = 500$ iterations, with a maximum of 512 tokens per response and a generation temperature of 0.8 to allow moderate exploration.

D.2 Quality comparisons of model outputs

Table 4 shows the responses generated by different methods using Llama-3-8B Instruct (Grattafiori et al., 2024) as the base model for the upper-level OpenOrca dataset. The responses generated by online self-refining are shorter yet clearer, which is consistent with the better performance on the tidy lower-level dataset. It also follows instructions more faithfully: for example, only online self-refining correctly rates the restaurant as $3/5$ rather than five stars. Moreover, while offline BDS provides a partially correct answer to the third question, it fails to adhere to the given options.

Table 5 gives an overview of the top 10% and bottom 10% questions ranked by offline BDS and online self-refining. The online self-refining approach shows a clear trend of ranking easier questions higher at the beginning and processing harder ones later. We also present template responses for some of the questions, sorted by difficulty level, in Table 6, which shows that response length can partially indicate question difficulty.

Online selection learns from simple to hard questions. To further compare the learning behavior of offline selection and online self-refining, we analyze the top 10% and bottom 10% questions at each epoch. Table 5 shows that online self-refining tends to focus on easy questions during the first epoch and gradually progresses from simple to hard questions. The knowledge gained from simpler questions provides a foundation for subsequent adaptation to harder questions (Ma et al., 2025; Zhang et al., 2025; Lin et al., 2025). We also report response length as a partial measure of question difficulty in Figure 4, where longer responses indicate longer reasoning and thus harder questions; see Table 6. Online self-refining initially emphasizes shorter-response questions, whereas offline BDS continues to emphasize longer-response questions throughout training.

Examples of generated responses from LLAMA-3-8B model fine-tuned by different methods

Human: John was a terrible writer. To practice, his teacher suggest that he consider people he knows and do what? Options: - write novels - advertising firm - write letter - write notes - write poems Let's think now! Step-by-step reasoning:

LLAMA-3-8B (Grattafiori et al., 2024):	**Offline Selection**:	**Online Self-refining**:
If John is a terrible writ... (read more) Option: B - advertising firm. Explanation: The correct answer is option B - "advertising firm". This is because the question states that John's teacher suggests that he "consider people he know" and "do what". In an advertising firm, one of the tasks would be to create ads for products or services, which requires writing. By doing so, John can practice his writing skills without having to write long, elaborate pieces like novels or poems. He can start with simple texts like advertisements, which are shorter and more focused on conveying information.	1. John is a terrible at writing. 2. His teacher suggests practicing to improve. 3. To practice writing, he can think about people he know. The answer would be "write letters". Because it's easier for him to start by thinking of someone he knows (like family or friends) and then writing a letter to them. This way he can practice his writing skills while also being more comfortable with the topic. Now let's see if we're right: Write your own explanation below in the space provided.	1. John is a terrible at writing. 2. His teacher suggests doing something to improve writing skills. 3. To improve writing, one can focus on writing about familiar topics or people they know. So the best answer is option C) write letters, as this will allow him to practice writing by thinking of people he know and writing a letter to them. This can be an easy way for him to get started with improving his writing.

Human: Generate an approximately fifteen-word sentence that describes all this data: Midsummer House eatType restaurant; Midsummer House food Chinese; Midsummer House priceRange moderate; Midsummer House customer rating 3 out of 5; Midsummer House near All Bar One

LLAMA-3-8B (Grattafiori et al., 2024):	**Offline Selection**:	**Online Self-refining**:
Midsummer House is a moderately priced Chinese restaurant with a 3-star rating nearby major hotels.	Midsummer House is a moderately priced, three-star-rated Chinese eatery located near All-Bar-One.	Midsummer House, a moderately-priced Chinese restaurant with a 3/5 rating, is located near All bar one.

Human: Premise: A man is inline skating in front of a wooden bench.
Hypothesis: A man is having fun skating in front of a bench.
Choose the correct answer: Given the premise, can we conclude the hypothesis?
Select from: a). yes b). it is not possible to tell c). no.

LLAMA-3-8B (Grattafiori et al., 2024):	**Offline Selection**:	**Online Self-refining**:
Correct answer: a) yes	d). maybe	b). it is not possible to tell

Table 4: Examples of generated response on OPENORCA dataset using LLAMA-3-8B (Grattafiori et al., 2024) and fine-tuned with offline BDS and online self-refining. Text marked in red indicates incorrect outputs, orange indicates partially correct outputs or irrelevant information, and green indicates fully correct outputs that match the expected instructions.

Method	Rank	Epoch 1	Epoch 3
Offline	Top	Classify the following text as either satire or non-satire. "The last presidential election was a great affair with exciting twists and turns that kept us all on our toes." [Medium] Generate a formal invitation for a networking event. We invite you to join us for an informal networking event. [Hard] Use a variety of language and words to rewrite the given sentence. He was very tired and cann't go any further. [Medium] Rewrite this code in C++. public static boolean isAnagram(String str1, String str2) char[] charArray1 = str1.toCharArray(); char[] charArray2 = str2.toCharArray(); Arrays.sort(charArray1); Arrays.sort(charArray2); return Arrays.equals(charArray1, charArray2); [Hard]	Find information about the primary schools in Johannesburg [Hard] Create a haiku poem using the provided words. Wind, Clouds, Sky [Hard] Identify the cause of this issue. The computer is not working. [Hard] Write a story using the given words in your story. desert, moonlit, violin. [Hard] Calculate the average speed of a car traveling 120 miles in 2 hours. 120 miles in 2 hours. [Medium]
	Bottom	Delete all words with more than 5 letters from this sentence. This sentence has many long words like 'sentence' and 'instruction'. [Easy] Rearrange the words in the sentence to form a question. Reading is difficult. [Easy] Is the following sentence structured correctly? We went for a walk in the park and played hide and seek. [Easy] Convert the following number in scientific notation: 0.567. [Easy]	Provide a list of materials needed for the given project. A school project to build a model of a volcano. [Hard] Is the following sentence structured correctly? We went for a walk in the park and played hide and seek. [Easy] Translate the following sentence from English to Spanish. Output less than 25 words. I am learning Spanish. [Easy] Write a code that prints the following string in all lowercase letters. Mixed Case Characters [Hard]
Online	Top	Find and replace all instances of the word "great" in the sentence with synonyms. The teacher's great approach in teaching helped the students to understand the lessons better. [Easy] Re-write the following sentence omitting the word "comfortable". We were quite comfortable with our decision. [Easy] Evaluate the following sentence and provide feedback on the spelling and punctuation errors. The frog jumpted acros the road. [Easy] Combine the sentences below into an essay. There are many ways to reduce waste. For example, reducing the use of plastic. Reusing materials is also important. [Easy]	Create a short story in the horror genre based on the given setting. Setting: An abandoned island [Hard] Given a set of data points, create an equation for the linear regression line. Data points: (1,1), (2,2), (3,4) [Hard] Come up with an experiment that tests the given concept. The effect of heavy metals on plant growth [Hard] Edit the given text such that its length is not changed but it clearly conveys the meaning. The app has got a very good user interface and is really nice to use. [Hard]
	Bottom	Write a script for a movie scene with the following input. Setting: Airport Characters: Rebecca and Jacob [Hard] Pick a random song and create a dance routine for it. "Just a Dream" by Nelly. [Hard] Write an email to a customer thanking them for their purchase. Customer Name – John Smith Product Purchased – Nike Shoes Delivery Date – 6/20/2021. [Hard] Explain the meaning of the following phrase. "out of the box". [Medium]	Assign the following verb a category of either transitive, intransitive, or linking. Jumped [Easy] Add a transition sentence at the end of this passage: Jack was so excited for his birthday party. He had spent months planning it. He had invited all his friends and family members. [Easy] Reverse the order of words in the sentence. Alice visited the museum. [Easy] Delete all words with more than 5 letters from this sentence. This sentence has many long words like 'sentence' and 'instruction'. [Easy]

Table 5: Examples of top 10% questions ranked by offline BDS and online self-refining. We annotate each question as [Easy], [Medium] or [Hard] based on whether they need creativity and deep thinking or just data analysis and reorganization. Online self-refining presents a clear trend of learning from simple to hard questions, while offline BDS tends to keep emphasizing harder questions. Learning from simple questions like 'combine the sentences below into an essay' serves as a good foundation for harder creativity questions like movie-script and poem writing.

Level	Question	Responses
Simple	Delete all words with more than 5 letters from this sentence. This sentence has many long words like 'sentence' and 'instruction'.	This has long like and.
Simple	Assign the following verb a category of either transitive, intransitive, or linking. Jumped.	Jumped is an intransitive verb.
Medium	Classify the following text as either satire or non-satire. "The last presidential election was a great affair with exciting twists and turns that kept us all on our toes."	This text would classify as non-satire.
Hard	Find information about the primary schools in Johannesburg.	Johannesburg is the largest city in South Africa and is home to a number of primary schools. These schools cater to students from the ages of 5-12 and provide education from Grade R (the Reception year) to Grade 7. Johannesburg's primary schools are either public (government-funded) or independent (privately-funded) institutions, and they offer the National Curriculum or the independent schools' curriculum. Some well-known primary schools in Johannesburg include: - Parkview Senior Primary School: A public school located in Parkview, Johannesburg. - Greenside Primary School: A public school located in Greenside, Johannesburg. - St Katharine's School : A private Anglican school located in Parktown, Johannesburg. - The Ridge School: An independent boys' school located in Westcliff, Johannesburg. - Pridwin Preparatory School: A private boys' school located in Melrose, Johannesburg. - Auckland Park Preparatory School: A private girls' school located in Auckland Park, Johannesburg. Admission requirements vary depending on the school, but most schools require an application form, an interview, and an assessment of the child's abilities. Potential students may also be placed on a waiting list. It is recommended to research the individual schools in Johannesburg and their specific admission requirements, fees, and curricula before making a decision on which school to enroll a child.
Hard	Write a script for a movie scene with the follow-ing input. Setting: Airport Characters: Rebecca and Jacob	Fade in: EXT. AIRPORT - DAY We open on a wide shot of a busy airport. People are rushing about with their luggage, trying to catch their flights. Amidst the hustle and bustle, we see REBECCA, mid-30s, standing with her arms crossed and tapping her foot impatiently. A moment later, JACOB, late-30s, appears, out of breath. JACOB: (Out of breath) Sorry, sorry. The traffic was insane. REBECCA: You're lucky I didn't leave without you. JACOB: (Smiling) You wouldn't do that to me. REBECCA: (Rolling her eyes) Don't be so sure. Jacob chuckles and the two make their way to the check-in counter. JACOB: Are you excited? REBECCA: (Sighs) I don't know. I'm nervous. JACOB: It'll be fine. You'll see. REBECCA: (Smiling) I'm glad you're here. JACOB: (Playfully nudging her) Always. The two check-in their luggage and make their way to their gate. REBECCA: Do you have the tickets? JACOB: (Pats his pockets, looking worried) Wait, I thought you had them. REBECCA looks at him sternly, but then bursts out laughing. REBECCA: I'm kidding. I have them. (Pulls out the tickets) JACOB: (Shakes his head, smiling) You're terrible. Their flight is called and the two join the line to board the plane, ready for their next adventure. Fade to black.

Table 6: Corresponding responses to part of questions in Table 5. Harder questions tend to have longer responses than simple and medium questions, but the distinction for simple and medium questions is not clear.

D.3 PER-BENCHMARK OOD QA EVALUATION

Tables 7 and 8 report zero-shot accuracy (%) for LLAMA-3-8B-INSTRUCT on the five OOD QA benchmarks. Each entry is the mean ± sample standard deviation over four independent training runs. For the Average column, we first compute the unweighted mean accuracy across the five benchmarks within each run, then summarize these four run averages. Thus, its standard deviation measures training-run variability.

Method	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-Easy	Average
Validation-only	81.38±0.81	75.13±0.83	64.85±0.54	66.05±0.79	75.63±0.61	72.61±0.43
Direct mixing ($\rho = 0.5$)	83.75±1.01	78.58±0.78	68.38±0.86	68.88±0.40	79.02±1.66	75.72±0.61
Direct mixing ($\rho = 1$)	85.28±0.39	79.65±0.34	68.83±0.70	66.74±0.59	79.45±0.56	75.99±0.37
Random selection	85.00±0.76	78.73±0.81	54.95±0.70	66.48±1.25	79.16±1.21	72.86±0.39
LESS	85.35±0.57	79.28±0.36	55.53±0.32	66.18±0.29	80.31±1.40	73.33±0.12
DSIR	85.58±0.32	79.13±0.49	54.15±0.71	68.40±0.18	79.36±0.33	73.32±0.22
Offline BDS	85.90±0.34	79.38±0.35	68.48±0.39	69.88±0.40	79.56±0.73	76.64±0.42
online selection	**87.28**±0.36	**81.18**±0.98	**69.18**±0.70	**70.28**±0.17	**81.81**±0.48	**77.94**±0.43

Table 7: Per-benchmark OOD QA accuracy (%) for the **quality-enhancement task**. Values are mean ± sample SD over four independent training runs. Higher is better; bold indicates the highest mean in each column.

Method	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-Easy	Average
Validation-only	82.03±0.48	75.45±0.54	62.73±0.40	65.10±0.47	77.13±0.68	72.49±0.26
Direct mixing ($\rho = 0.5$)	83.50±0.55	75.63±0.40	63.33±0.46	65.88±0.35	76.71±1.97	73.01±0.28
Direct mixing ($\rho = 1$)	81.70±0.32	75.65±0.39	62.78±0.33	62.60±0.39	74.36±2.07	71.42±0.58
Random selection	83.95±0.62	77.70±0.39	63.68±0.43	67.80±0.48	78.30±0.89	74.29±0.36
LESS	84.40±0.56	78.05±0.39	63.80±0.37	68.23±0.38	76.85±0.81	74.27±0.35
DSIR	84.30±0.47	77.75±0.42	63.73±0.49	67.98±0.40	78.85±1.88	74.52±0.30
Offline BDS	84.25±0.47	78.03±0.22	66.90±0.37	68.95±0.62	**79.84**±1.49	75.59±0.37
Online dynamic	**85.63**±0.40	**78.75**±0.39	**68.40**±0.50	**71.45**±0.52	78.54±0.69	**76.55**±0.34

Table 8: Per-benchmark OOD QA accuracy (%) for the **safety-aware task** with BLUEORCA guidance. Values are mean ± sample SD over four independent training runs. Higher is better; bold indicates the highest mean in each column.

D.4 COMPONENT-WISE ABLATION

Table 9 separates question selection, response regeneration, and BMO reweighting on the quality-enhancement task with LLAMA-3-8B-INSTRUCT. Direct mixing retains all lower-level questions, whereas Offline BDS applies question selection. The equal-weight baseline adds regenerated responses without BMO weighting; both regenerated-response methods use $R = 10\%$ and $G = 1, 5$.

Method	G	Question selection	Online response regeneration	Importance-ratio weighting	OpenOrca loss ↓
Direct mixing ($\rho = 0.5$)	–	No	No	No	0.84 ± 0.008
Offline BDS	–	Yes	No	No	0.80 ± 0.007
Equal-weight regeneration	1	Yes	Yes	No	0.82 ± 0.005
Online BDS	1	Yes	Yes	Yes	0.78 ± 0.006
Equal-weight regeneration	5	Yes	Yes	No	0.81 ± 0.004
Online BDS	5	Yes	Yes	Yes	$\mathbf{0.76} \pm 0.003$

Table 9: Component-wise comparison on the quality-enhancement task. Values are held-out OpenOrca loss, reported as mean ± SD. Both regeneration variants use $R = 10\%$; G denotes the number of generated responses per question.

Question selection reduces the mean loss, but equal-weight regeneration gives slightly higher loss than the offline BDS, showing that adding online model-generated responses does not by itself improve the offline BDS. Importance-ratio reweighting is important to reduce enhance the online data selection performance, supporting its contribution to online refinement.

D.5 HARMFULNESS EVALUATION

Following the harmful-instruction evaluation of Qi et al. (2024), we evaluate the generated responses on 300 held-out HEx-PHI prompts. The methods use the same evaluation prompts and generation settings. Each response is scored once by GPT-4.1-mini on a 1–5 harmfulness scale, where lower is safer. Figure 3(d) reports the arithmetic mean of these scores.

D.6 IMPORTANCE-RATIO APPROXIMATED RESPONSE-LEVEL WEIGHTS

Importance-ratio versus response-level BMO weights. In a paired LLAMA-3-8B experiment, we sample $G = 5$ responses for each of 512 REDORCA questions and perform a 500-step online refresh using BLUEORCA as the upper-level data. We compare importance-ratio coefficients with the exact response-level BMO reference weights. Table 10 summarizes the average cosine similarity across questions and iterations.

Weight-agreement metric	Mean
Cosine similarity ↑	0.984 ± 0.003

Table 10: Agreement between importance-ratio and response-level BMO weights.

LIMITATIONS AND BORADER IMPACTS

Our theoretical guarantees rely on the separable data assumption (Assumption 1), which requires the overparameterized SFT model to interpolate the training data. In practice, without this assumption, one may need to run additional gradient descent on the lower-level objective to estimate its minimal function value and subtract it from the penalty objective (Shen et al., 2025b). Our method may benefit the society by improving the quality and safety of SFT data used for LLM.