RETHINKING INFLUENCE-BASED DATA SELECTION: A CURVATURE-AWARE SAMPLING FRAMEWORK FOR VISUAL INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting informative training data is central to efficient multimodal instruction tuning. While existing methods widely rank examples via first-order, pointwise influence on validation loss, such filtering provides no guarantees on the validation risk after subset retraining. To address this limitation, we introduce a budget-constrained Poisson sampling framework that optimizes inclusion probabilities to minimize the expected validation risk after subset retraining. Crucially, we show that the validation curvature governs the retrained risk and derive a second-order risk surrogate endowed with a closed form optimal inclusion rule. We further establish an excess risk bound that decays as $o(B^{-1})$ with the sampling budget B, and characterize the relationship between existing influence function methods and the proposed framework. Extensive experiments across multiple vision language models, multimodal benchmarks, and selection budgets demonstrate the effectiveness of our method; notably, with an expected selection budget of only 6.4%, it retains 98.38% of the performance of training on full dataset.

1 INTRODUCTION

Large-scale visual instruction tuning (VIT) (Liu et al., 2024a; 2023; Zhao et al., 2023) has become a cornerstone for advancing the capabilities of modern vision language models (VLMs) (Ghosh et al., 2024; Li et al., 2025; Zhang et al., 2024). By leveraging massive collections of image instruction pairs, recent VLMs have demonstrated strong performance across a wide range of multimodal tasks, such as visual question answering, diagram understanding, and multimodal reasoning (Lian et al., 2025; Liu et al., 2024b; 2026; Wei et al., 2025; Wu et al., 2025; 2024). However, as visual instruction datasets continue to grow, concerns about training efficiency and data effectiveness have become increasingly prominent. Training on all available samples can even underperform training on a selected subset (Xia et al., 2024; Wu et al., 2026).

A key direction for addressing the challenges of training efficiency and data effectiveness is data selection (Ting & Brochu, 2018; Koh & Liang, 2017; Xia et al., 2024; Nikdan et al., 2025; Xu et al., 2026; Wu et al., 2026), which aims to select training examples that reduce downstream task risk. There are two widely used approaches to data selection: influence functions (Kwon et al., 2024) and first-order validation risk gradient based methods (Xia et al., 2024). Both approaches estimate each candidate example's influence on validation risk and use the resulting pointwise scores to rank or filter training examples. However, selecting examples according to these scores does not guarantee that retraining on the selected subset minimizes validation risk under a budget constraint.

Our contributions To address this limitation, we formulate budget constrained data selection as the optimization of inclusion probabilities to minimize expected validation risk after subset retraining. Under Poisson sampling framework, we combine a linearization of the weighted training-loss gradient with a second-order expansion of validation risk to derive a curvature-aware risk surrogate with a closed-form optimal inclusion design, which we term Curvature-Aware Influence Sampling (CAIS). Furthermore, we establish an $o(B^{-1})$ bound on the excess expected validation risk of surrogate sampling design relative to the optimal sampling strategy, where B denotes the expected sampling budget. Within our framework, classi-

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.

cal influence-based sampling (Koh & Liang, 2017; Ting & Brochu, 2018) can be derived and interpreted as variance minimization under a first-order approximation of the validation risk. Consequently, it relies solely on the rank-one outer product of the validation gradient, whereas CAIS leverages the validation curvature. To enable sampling design computationally tractable, we use gradient sketching to approximate the inclusion probability in a low-dimensional space. Experiments across multiple vision language models, multimodal benchmarks, and selection budgets demonstrate the effectiveness of our approach. As summarized in Figure 1, it outperforms the strongest selection baseline by 2.41 and 1.90 percentage points under the 40K and 100K budgets.

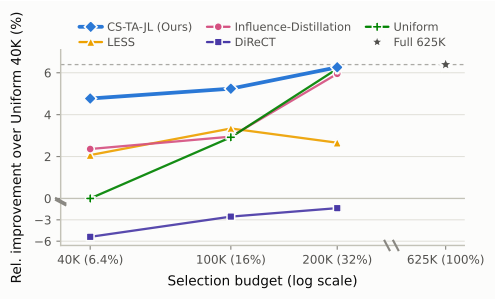

Figure 1: Downstream performance of LLaVA-Vicuna-7B on LLaVA-V-625K selected subsets.

2 Preliminaries

Let $\mathcal{D}_{\text{tr}} = \{Z_i\}_{i=1}^{n}$ and $\mathcal{D}_{\text{val}} = \{Z_j^{\text{val}}\}_{j=1}^{m}$ denote the training and validation sets, respectively. Each example $Z_i = (V_i, X_i, Y_i)$ consists of a visual input V_i, a textual prompt X_i, and a target response Y_i. For a vision language model p_θ with trainable parameters $\theta \in \mathbb{R}^d$, the loss of Z_i is

$$\ell(Z_i;\theta) := -\frac{1}{T_i}\sum_{t=1}^{T_i} \log p_\theta(Y_{i,t} \mid V_i, X_i, Y_{i,<t}), \tag{1}$$

where T_i is the number of supervised response tokens. The training and validation risks are

$$F_{\text{tr}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i;\theta), \qquad F_{\text{val}}(\theta) := \frac{1}{m}\sum_{j=1}^{m} \ell(Z_j^{\text{val}};\theta). \tag{2}$$

Let $\hat{\theta}$ denote an empirical risk minimizer of the training risk, i.e., $\hat{\theta} \in \arg\min_\theta F_{\text{tr}}(\theta)$. We write $H_{\text{tr}}(\theta) := \nabla_\theta^2 F_{\text{tr}}(\theta)$ for the training risk Hessian and define $H_{\text{val}}(\theta)$ analogously.

2.1 Influence Functions

Influence functions measure the local sensitivity of model parameters to the weight assigned to an individual training example (Koh & Liang, 2017; Kwon et al., 2024). Specifically, upweighting Z_i by some small ϵ gives us new parameters $\hat{\theta}_{\epsilon,i} := \arg\min_\theta\{F_{\text{tr}}(\theta) + \epsilon\ell(Z_i;\theta)\}$. Assume that the empirical risk is twice differentiable and strictly convex in θ, with a positive definite Hessian at $\hat{\theta}$. Following Cook & Weisberg (1982), the influence function is

$$\mathcal{I}_{\hat{\theta}}(Z_i) := \left.\frac{\mathrm{d}\hat{\theta}_{\epsilon,i}}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = -H_{\text{tr}}(\hat{\theta})^{-1}\nabla_\theta \ell(Z_i;\hat{\theta}). \tag{3}$$

The influence of Z_i on validation loss follows from the chain rule (Koh & Liang, 2017):

$$\begin{aligned}\mathcal{I}(Z_i) &:= \left.\frac{\mathrm{d}F_{\text{val}}(\hat{\theta}_{\epsilon,i})}{\mathrm{d}\epsilon}\right|_{\epsilon=0} \\ &= -\nabla_\theta F_{\text{val}}(\hat{\theta})^\top H_{\text{tr}}(\hat{\theta})^{-1}\nabla_\theta \ell(Z_i;\hat{\theta}).\end{aligned} \tag{4}$$

A positive $\mathcal{I}(Z_i)$ means that infinitesimally upweighting Z_i locally increases validation loss, whereas a negative value indicates a local decrease.

2.2 Poisson Sampling Design

A Poisson design (Ting & Brochu, 2018) draws independent indicators $W_i \sim \text{Bernoulli}(\pi_i)$, where $\Pr(W_i = 1) = \pi_i$ and $\Pr(W_i = 0) = 1 - \pi_i$, with $W_i = 1$ indicating that example Z_i is selected.

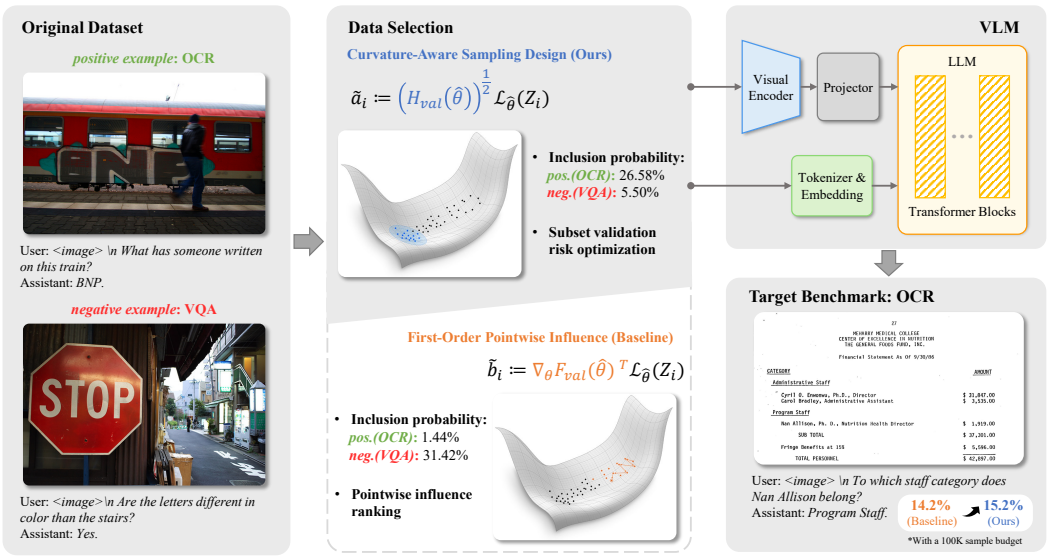

Figure 2: Two candidates scored under both designs, with OCR as the downstream task. Our curvature-aware design gives the text-recognition example a 26.58% inclusion probability against 1.44% under first-order pointwise influence, which instead favors a task mismatched color comparison example (31.42% versus 5.50%) and scores lower on OCRBench under a 100K budget.

All expectations $\mathbb{E}_W$ below are over the sampling indicators. The number of selected examples is $\sum_{i=1}^{n} W_i$, with expectation

$$\mathbb{E}_W\left[\sum_{i=1}^{n} W_i\right] = \sum_{i=1}^{n} \pi_i. \tag{5}$$

We specify a budget $B > 0$ as an upper bound on the expected number of selected examples and impose a lower bound $\alpha \in (0, 1)$ on each inclusion probability. The feasible designs are

$$\mathcal{P}_\alpha(B) := \left\{ \pi \in [\alpha, 1]^n : \sum_{i=1}^{n} \pi_i \le B \right\}. \tag{6}$$

To mitigate selection bias and obtain an unbiased estimate of training objective, we follow Ting & Brochu (2018) and employ the Horvitz–Thompson (HT) weighted training loss:

$$F_{\mathrm{tr},\pi}(\theta; W) := \frac{1}{n} \sum_{i=1}^{n} \frac{W_i}{\pi_i} \ell(Z_i; \theta). \tag{7}$$

Since $\mathbb{E}_W[W_i/\pi_i] = 1$, we have $\mathbb{E}_W[F_{\mathrm{tr},\pi}(\theta; W)] = F_{\mathrm{tr}}(\theta)$. Let $\hat{\theta}_\pi(W) \in \arg\min_\theta F_{\mathrm{tr},\pi}(\theta; W)$, assuming such a solution exists. The lower bound α is imposed to avoid excessively large HT weights and high estimation variance of $\hat{\theta}_\pi(W)$.

3 Method

We formulate data selection as the design of inclusion probabilities under a budget. Our objective is the expected validation risk after Poisson sampling and HT-weighted retraining. We derive a tractable criterion by combining a linearization of the weighted training risk gradient with a second-order expansion of validation risk, then show how the same framework connects classical influence based sampling through first-order variance minimization. Figure 2 illustrates the gap on an OCR target task, where the two designs disagree on which candidates to sample. Throughout, the training and validation dataset are treated as fixed, so all randomness comes from the sampling indicators.

3.1 Expected Validation Risk under Subsampling

We choose inclusion probabilities to minimize the expected validation risk after HT-weighted retraining:

$$\min_{\pi \in \mathcal{P}_\alpha(B)} \mathcal{R}(\pi), \qquad \mathcal{R}(\pi) := \mathbb{E}_W \left[F_{\mathrm{val}}(\hat{\theta}_\pi(W)) \right]. \tag{8}$$

Directly evaluating this empirical objective requires repeated retraining for each candidate design. We obtain a tractable local surrogate by combining the influence function with the curvature of the validation risk.

Influence approximation of retraining A first-order Taylor expansion around $\hat{\theta}$ gives

$$\begin{aligned} \nabla_\theta F_{\mathrm{tr},\pi}(\hat{\theta}_\pi(W); W) &\approx \nabla_\theta F_{\mathrm{tr},\pi}(\hat{\theta}; W) + \nabla^2_\theta F_{\mathrm{tr},\pi}(\hat{\theta}; W)\big(\hat{\theta}_\pi(W) - \hat{\theta}\big), \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \left(\frac{W_i}{\pi_i} - 1 \right) \nabla_\theta \ell(Z_i; \hat{\theta}) + H_{\mathrm{tr}}(\hat{\theta})\big(\hat{\theta}_\pi(W) - \hat{\theta}\big), \end{aligned} \tag{9}$$

where the first term follows from the definition of the HT-weighted training risk in Eq. 7 together with $\nabla_\theta F_{\mathrm{tr}}(\hat{\theta}) = 0$, and the second term follows from replacing the sampled Hessian by $H_{\mathrm{tr}}(\hat{\theta})$. Since the left-hand side of Eq. 9 is zero, we have

$$\begin{aligned} \hat{\theta}_\pi(W) - \hat{\theta} &\approx -H_{\mathrm{tr}}(\hat{\theta})^{-1} \frac{1}{n}\sum_{i=1}^{n} \left(\frac{W_i}{\pi_i} - 1 \right) \nabla_\theta \ell(Z_i; \hat{\theta}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{W_i}{\pi_i} - 1 \right) \mathcal{I}_{\hat{\theta}}(Z_i). \end{aligned} \tag{10}$$

Note that Eq. 10 has zero mean, which eliminates the expected first-order term in the validation risk expansion, leaving a quadratic term from which we construct the surrogate.

A curvature-aware validation risk surrogate Similarly, a second-order Taylor expansion of the validation risk around $\hat{\theta}$ gives

$$\begin{aligned} F_{\mathrm{val}}(\hat{\theta}_\pi(W)) \approx F_{\mathrm{val}}(\hat{\theta}) &+ \nabla_\theta F_{\mathrm{val}}(\hat{\theta})^\top \big(\hat{\theta}_\pi(W) - \hat{\theta}\big) \\ &+ \frac{1}{2}\big(\hat{\theta}_\pi(W) - \hat{\theta}\big)^\top H_{\mathrm{val}}(\hat{\theta})\big(\hat{\theta}_\pi(W) - \hat{\theta}\big). \end{aligned} \tag{11}$$

Substituting $\hat{\theta}_\pi(W) - \hat{\theta}$ from Eq. 10 into the linear and quadratic terms of Eq. 11, then taking expectations over W, defines the surrogate

$$\widetilde{\mathcal{R}}(\pi) := F_{\mathrm{val}}(\hat{\theta}) + \frac{1}{2n^2}\sum_{i=1}^{n} \left(\frac{1}{\pi_i} - 1 \right) a_i. \tag{12}$$

The coefficient $a_i := \mathcal{I}_{\hat{\theta}}(Z_i)^\top H_{\mathrm{val}}(\hat{\theta}) \mathcal{I}_{\hat{\theta}}(Z_i)$ measures validation curvature along the influence function of example Z_i. For nonnegative coefficients, increasing π_i reduces this example's contribution to the surrogate; larger a_i therefore favor larger inclusion probabilities under the budget constraint. Suppose $H_{\mathrm{val}}(\hat{\theta})$ is positive semidefinite, so that $a_i \geq 0$. Dropping terms independent of π from the quadratic surrogate in Eq. 12 gives

$$\min_{\pi} \quad \sum_{i=1}^{n} \frac{a_i}{\pi_i}, \quad \text{subject to} \quad \sum_{i=1}^{n} \pi_i \leq B. \tag{13}$$

The standard Karush–Kuhn–Tucker conditions (Boyd & Vandenberghe, 2004) yield

$$\pi^\star_{\alpha,i} = \max\{\alpha, \min(1, \kappa\sqrt{a_i})\}, \qquad \sum_{i=1}^{n} \pi^\star_{\alpha,i} = B, \tag{14}$$

where $\kappa > 0$ is chosen to satisfy the budget.

3.2 Validation Excess Risk Bound

In the last section, we established an optimal sampling rule for a quadratic surrogate of the expected validation risk. To assess its expected validation risk after retraining, we compare it with the optimal sampling design for the original objective under the same constraints:

$$\pi_{\mathrm{opt}} \in \operatorname*{arg\,min}_{\pi \in \mathcal{P}_\alpha(B)} \mathcal{R}(\pi). \tag{15}$$

The following theorem bounds the excess expected validation risk of $\pi_\alpha^\star$ relative to the optimal design π_{opt}.

Theorem 1 (Nearly optimal surrogate design.)**.** *Consider a sequence of datasets indexed by the budget $B \to \infty$, with $B \le n$. Suppose the conditions in Appendix A.1 hold separately for $\pi = \pi_\alpha^\star$ and $\pi = \pi_{\mathrm{opt}}$, and that $\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2 = O(r_B)$ for some deterministic sequence $r_B > 0$. Then*

$$0 \le \mathcal{R}(\pi_\alpha^\star) - \mathcal{R}(\pi_{\mathrm{opt}}) = O\Big(\frac{r_B}{B} + B^{-3/2}\Big).$$

In particular, the excess risk is $o(B^{-1})$ when $r_B = o(1)$, and $O(B^{-3/2})$ when $r_B = O(B^{-1/2})$.

Appendix A.2 proves the theorem and extends the analysis to a fixed, possibly nonstationary checkpoint of full data training risk and the HT-weighted training risk. When $r_B = o(1)$, we further prove in Proposition 5 of Appendix A.3 that the excess validation risk of $\pi_\alpha^\star$ relative to π_{opt} decays asymptotically faster than that of uniform sampling. Appendix A.4 shows that these conditions are not vacuous: they hold for regularized frozen feature softmax models, and in a binary cross-entropy example the validation gradient converges to zero in probability as both sample sizes grow, so that $r_B \to 0$.

3.3 Connection to Influence Function Based Sampling

Consider the first-order approximation of the validation risk:

$$F_{\mathrm{val}}(\hat{\theta}_\pi(W)) \approx F_{\mathrm{val}}(\hat{\theta}) + \nabla_\theta F_{\mathrm{val}}(\hat{\theta})^\top \big(\hat{\theta}_\pi(W) - \hat{\theta}\big). \tag{16}$$

Substituting $\hat{\theta}_\pi(W) - \hat{\theta}$ from Eq. 10 then gives

$$F_{\mathrm{val}}(\hat{\theta}_\pi(W)) \approx F_{\mathrm{val}}(\hat{\theta}) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\pi_i} - 1\right)\mathcal{I}(Z_i), \tag{17}$$

where $\mathcal{I}(Z_i)$ is the validation risk influence in Eq. 4. Taking expectations over W on both sides of Eq. 17, the random correction on the right has zero mean because $\mathbb{E}_W[W_i/\pi_i - 1] = 0$. The expected validation risk is therefore approximately $F_{\mathrm{val}}(\hat{\theta})$ within this first-order approximation, regardless of the sampling design. This naturally motivates minimizing the variance of the correction to reduce validation risk fluctuations across sampled subsets. Independence of the sampling indicators gives

$$\mathrm{Var}_W\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\pi_i} - 1\right)\mathcal{I}(Z_i)\right] = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{\pi_i} - 1\right) b_i, \tag{18}$$

where

$$b_i := \mathcal{I}(Z_i)^2 = \mathcal{I}_{\hat{\theta}}(Z_i)^\top \big[\nabla_\theta F_{\mathrm{val}}(\hat{\theta}) \nabla_\theta F_{\mathrm{val}}(\hat{\theta})^\top\big] \mathcal{I}_{\hat{\theta}}(Z_i). \tag{19}$$

Minimizing this variance under the expected budget reduces to $\min_{\pi \in \mathcal{P}_\alpha(B)} \sum_i b_i/\pi_i$. Following the solution in Eq. 14, for $b_i > 0$ and $0 < B < n$, we obtain

$$\pi_{\alpha,i}^{\mathrm{IF}} = \max\Big\{\alpha, \min(1, \kappa_{\mathrm{IF}}\sqrt{b_i})\Big\}, \qquad \sum_{i=1}^{n} \pi_{\alpha,i}^{\mathrm{IF}} = B. \tag{20}$$

Before saturation, the probabilities are proportional to $\sqrt{b_i} = |\mathcal{I}(Z_i)|$. This coincides with the optimal influence function based subsampling rule of Ting & Brochu (2018). Compared with a_i in Eq. 12, b_i uses the validation gradient outer product in place of the validation Hessian. Consequently, a influence function orthogonal to the validation gradient has $b_i = 0$ but can still have $a_i > 0$.

3.4 Efficient Computation of the Sampling Design

Let d denote the number of trainable parameters, with Hessian $H \in \mathbb{R}^{d\times d}$. For large language models ($d \sim 10^9$), explicitly forming and inverting H is computationally prohibitive. We approximate the coefficients a_i using a blockwise version of CountSketch (Charikar et al., 2002) for gradient estimation and Gauss Newton Hessian (GNH) estimates (Grosse et al., 2023) for curvature.

We first apply a Blockwise CountSketch map $S \in \mathbb{R}^{k\times d}$, where $k \ll d$ is the sketch dimension, to obtain $S\nabla_\theta \ell(Z_i;\hat{\theta}) \in \mathbb{R}^k$ for each candidate. Following Grosse et al. (2023), we estimate the GNH using L independent pseudo-label draws $\widetilde{Y}_{i,t}^{(l)} \sim p_{\hat{\theta}}(\cdot \mid V_i, X_i, Y_{i,<t})$, $l = 1,\dots,L$, at each supervised position. We then compute

$$
\begin{aligned}
h_i^{(l)} &:= -\frac{1}{\sqrt{T_i}} \sum_{t=1}^{T_i} \nabla_\theta \log p_\theta(\widetilde{Y}_{i,t}^{(l)} \mid V_i, X_i, Y_{i,<t})\Big|_{\theta=\hat{\theta}}, \\
\widehat{H}_{\mathrm{tr}} &:= \frac{1}{nL} \sum_{i=1}^{n} \sum_{l=1}^{L} (Sh_i^{(l)})(Sh_i^{(l)})^\top \in \mathbb{R}^{k\times k}.
\end{aligned}
\tag{21}
$$

The same estimator applies to validation data, replacing n by m, and gives $\widehat{H}_{\mathrm{val}} \in \mathbb{R}^{k\times k}$. Writing $x_i := S\nabla_\theta \ell(Z_i;\hat{\theta})$, we approximate $\mathcal{I}_{\hat{\theta}}(Z_i)$ in the sketched space by $-(\widehat{H}_{\mathrm{tr}} + \zeta I_k)^{-1} x_i$ with damping $\zeta > 0$, and compute

$$
\widehat{a}_i := x_i^\top (\widehat{H}_{\mathrm{tr}} + \zeta I_k)^{-1} \widehat{H}_{\mathrm{val}} (\widehat{H}_{\mathrm{tr}} + \zeta I_k)^{-1} x_i \tag{22}
$$

as the numerical approximation of a_i in Eq. 12. Since $\widehat{H}_{\mathrm{val}}$ is positive semidefinite, $\widehat{a}_i \geq 0$.

4 Experiments

4.1 Experimental Setup

Datasets We follow ScalSelect (Wu et al., 2026) in the choice of model families and evaluation benchmarks. Our candidate training set $\mathcal{D}_{\mathrm{tr}}$ is LLaVA-V-625K, obtained by excluding 40K text-only examples from LLaVA-665K (Liu et al., 2024a). The resulting 625K multimodal examples cover visual question answering, optical character recognition, region level question answering, and visual conversations. For data selection, we construct $\mathcal{D}_{\mathrm{val}}$ by randomly sampling 20% of the benchmark test examples within each task, reserving the remaining 80% as the test set.

Models and training We use LLaVA-Vicuna-7B as the main model, initialized from the LLaVA-1.5-7B checkpoint before visual instruction tuning (Liu et al., 2024a). Additional experiments use Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct (Bai et al., 2025) to examine performance across model families and sizes. All models are LoRA (Hu et al., 2022) fine-tuned for one epoch in main results. On LLaVA-V-625K, we consider selection budgets $B \in \{40\mathrm{K}, 100\mathrm{K}, 200\mathrm{K}\}$, corresponding to 6.4%, 16%, and 32% of the candidate set. For our sampling framework, B denotes the expected number of selected examples (possibly slightly more or fewer in practice), consistent with Section 2.2.

Evaluation benchmarks We report nine metrics across six multimodal benchmarks. MMBench (Liu et al., 2024c) evaluates general multimodal understanding in English (MMBench-En) and Chinese (MMBench-Cn), while MME (Fu et al., 2023) measures perception (MME-P) and cognition (MME-C). AI2D (Kembhavi et al., 2016) assesses diagram understanding. For object hallucination, we report accuracy (POPE-A) and precision (POPE-P) on POPE (Li et al., 2023). We further evaluate scientific question answering on the image subset of ScienceQA (SQA-IMG) (Lu et al., 2022) and text recognition and understanding on OCRBench (Liu et al., 2024d).

Baselines We compare with three methods that use gradient or curvature information for data selection. **LESS** (Xia et al., 2024) ranks candidate examples using cosine similarity between their Adam update features and validation gradients, aggregated across warm-up checkpoints; LoRA adaptation and random projection reduce the cost of storing and comparing these features. **Influence**

Distillation (Nikdan et al., 2025) derives sample weights from a second-order approximation of the target loss after a training update, and scales selection by computing influences on a small landmark set and propagating them to the remaining candidates. **DiReCT** (Xu et al., 2026) formulates selection as a constrained optimization problem using the spectral geometry of validation curvature: it favors progress along low curvature directions while limiting gradient components in high curvature directions. **Uniform** randomly samples from the candidate training set, assigning the same inclusion probability to every example. We also include **Full**, obtained by fine-tuning the model on the entire training set. "Rel." denotes the average relative performance across benchmarks, computed as the mean percentage normalized by the reference method set to 100% in each table.

4.2 Main Results

All results in this section are averaged over three repeated runs, and $\pm$ in the tables denotes the standard deviation across these runs. Tables 1, 2, and 6 (Appendix B) report the performance comparison with baselines on LLaVA-Vicuna-7B under the 40K, 100K, and 200K selection budgets, respectively. CAIS achieves the highest reported Rel. among the compared data selection methods under all three selection budgets. Under the 40K selection budget, corresponding to only 6.4% of the candidate training set, it reaches a Rel. of 98.38%, exceeding Uniform and the strongest selection baseline, Influence Distillation, by 4.77 and 2.41 percentage points, respectively. Under the 100K selection budget, CAIS reaches 98.85%, outperforming Uniform by 2.32 points and LESS, the strongest selection baseline at this budget, by 1.90 points. The larger margin over Uniform under the 40K selection budget suggests that incorporating validation curvature helps our method identify influential training examples more precisely, yielding greater benefits under tight selection budgets.

The aggregate gains are accompanied by improvements on several individual metrics. Under the 100K selection budget, CAIS exceeds Full on MMBench-En, MME-P, and OCRBench, although its MME-C score is lower. Thus, its strong aggregate performance reflects gains across multiple tasks rather than uniform improvements on every metric.

Table 1: Performance comparison with baselines under the 40K selection budget.

Method	MMBench-En	MMBench-Cn	MME-P	MME-C	AI2D	POPE-A	POPE-P	SQA-IMG	OCRBench	Rel.
Full	61.81±1.02	54.15±0.38	1319.64±5.60	268.75±3.78	54.01±1.56	83.40±1.31	93.72±1.89	69.39±0.78	14.70±0.53	100.00%
Uniform	55.53±1.30	49.36±0.60	1278.68±7.31	266.96±4.33	46.96±2.89	83.99±2.01	84.80±2.37	65.68±1.23	13.60±0.89	93.61%
Influence-Distillation	60.74±1.05	50.32±2.03	1234.41±7.03	282.59±2.99	50.85±1.32	83.08±0.42	84.16±0.69	67.84±0.61	13.60±0.49	95.97%
LESS	58.40±1.35	53.94±0.23	1202.56±6.85	273.21±4.74	50.81±1.95	81.13±4.48	76.97±6.60	67.53±0.26	15.20±0.85	95.68%
DiReCT	53.19±0.68	43.72±0.15	1229.90±14.66	232.14±3.47	41.42±2.20	80.60±4.12	94.83±2.48	55.89±5.17	13.60±0.35	88.22%
CAIS (Ours)	61.17±0.54	52.98±0.32	1211.01±4.21	297.32±7.51	51.82±3.67	84.35±2.31	87.45±1.77	68.84±0.57	14.20±0.39	98.38%

Table 2: Performance comparison with baselines under the 100K selection budget.

Method	MMBench-En	MMBench-Cn	MME-P	MME-C	AI2D	POPE-A	POPE-P	SQA-IMG	OCRBench	Rel.
Full	61.81±1.02	54.15±0.38	1319.64±5.60	268.75±3.78	54.01±1.56	83.40±1.31	93.72±1.89	69.39±0.78	14.70±0.53	100.00%
Uniform	56.28±1.02	52.77±0.47	1237.42±5.72	266.96±3.41	50.97±2.24	82.17±1.58	93.97±1.85	67.10±0.95	14.30±0.69	96.53%
Influence-Distillation	61.81±0.83	51.60±2.26	1240.01±6.41	258.93±3.16	49.84±0.20	84.19±2.31	93.45±0.59	67.04±0.26	13.80±0.64	96.56%
LESS	61.06±0.23	55.53±1.20	1202.61±11.81	291.07±2.83	51.13±0.06	83.64±5.86	80.79±9.04	66.17±0.79	14.00±0.42	96.95%
DiReCT	50.74±1.65	43.19±3.39	1317.39±10.70	301.79±2.67	40.53±0.49	81.13±0.88	94.45±1.50	51.24±0.79	14.50±0.21	91.06%
CAIS (Ours)	62.77±0.42	53.83±0.25	1340.83±3.29	231.70±5.86	53.81±2.86	82.93±1.80	92.49±1.38	69.21±0.44	15.20±0.30	98.85%

Table 3 reports the performance of our method across different model architectures under the 40K selection budget. The advantage over Uniform also appears on these models. Under the 40K selection budget, CAIS improves Rel. from 96.24% to 98.33% on Qwen3-VL-4B and from 96.97% to 98.51% on Qwen3-VL-8B, yielding gains of 2.09 and 1.54 percentage points, respectively. CAIS has higher mean scores on 7/9 (Qwen3-VL-4B) and 8/9 (Qwen3-VL-8B) metrics, including both MMBench variants and both MME metrics. These results extend the evidence beyond LLaVA and support the applicability of the sampling framework across model families and sizes.

4.3 Extended Results

Comparison with influence function based sampling Table 4 compares CAIS with Influence Function Based Sampling (IFS) under the three selection budgets. CAIS obtains higher Rel. scores

Table 3: Performance comparison with baselines across different model architectures under the 40K selection budget.

Model	Method	#Samples	MMBench-En	MMBench-Cn	MME-P	MME-C	AI2D	POPE-A	POPE-P	SQA-IMG	OCRBench	Rel.
Qwen3-VL-4B	Full	625K	$85.00_{\pm 0.93}$	$83.09_{\pm 0.41}$	$1676.24_{\pm 5.17}$	$634.82_{\pm 3.45}$	$83.28_{\pm 1.43}$	$89.75_{\pm 1.18}$	$96.87_{\pm 1.62}$	$91.82_{\pm 0.73}$	$57.80_{\pm 0.48}$	100.00%
	Uniform	40K	$81.60_{\pm 1.24}$	$79.15_{\pm 0.55}$	$1588.92_{\pm 6.89}$	$584.38_{\pm 4.17}$	$80.73_{\pm 2.58}$	$88.96_{\pm 1.83}$	$93.64_{\pm 2.08}$	$91.51_{\pm 1.11}$	$55.30_{\pm 0.82}$	96.24%
	CAIS	40K	$82.45_{\pm 0.49}$	$80.11_{\pm 0.29}$	$1660.07_{\pm 3.87}$	$637.95_{\pm 6.43}$	$80.61_{\pm 3.19}$	$89.35_{\pm 2.07}$	$94.75_{\pm 1.59}$	$91.26_{\pm 0.53}$	$56.90_{\pm 0.35}$	98.33%
Qwen3-VL-8B	Full	625K	$86.91_{\pm 0.85}$	$84.04_{\pm 0.35}$	$1682.28_{\pm 4.94}$	$683.48_{\pm 3.26}$	$85.06_{\pm 1.25}$	$88.68_{\pm 1.05}$	$97.01_{\pm 1.49}$	$93.93_{\pm 0.70}$	$61.00_{\pm 0.46}$	100.00%
	Uniform	40K	$83.83_{\pm 1.16}$	$80.43_{\pm 0.48}$	$1643.25_{\pm 6.76}$	$631.70_{\pm 3.92}$	$83.36_{\pm 2.37}$	$88.68_{\pm 1.64}$	$93.30_{\pm 1.93}$	$93.06_{\pm 1.02}$	$59.30_{\pm 0.77}$	96.97%
	CAIS	40K	$85.00_{\pm 0.45}$	$82.13_{\pm 0.26}$	$1687.55_{\pm 3.60}$	$669.64_{\pm 5.94}$	$83.16_{\pm 2.98}$	$89.39_{\pm 1.90}$	$93.32_{\pm 1.48}$	$93.91_{\pm 0.50}$	$59.80_{\pm 0.31}$	98.51%

than IFS under all three selection budgets. The largest gap occurs under the 40K selection budget, where CAIS improves all nine metrics and raises Rel. from 94.54% to 98.38%. Under the 100K and 200K selection budgets, the aggregate differences narrow to 0.29 and 0.40 percentage points, respectively, with mixed results across individual metrics. This pattern supports the motivation for incorporating validation curvature into the sampling design, with the clearest empirical advantage observed under the smallest selection budget.

To identify where the two designs disagree, we focus on the two data sources in the training set most relevant to OCR, OCR-VQA and TextVQA, and compare the subsets the two designs select from them under the 40K budget. Of the examples CAIS selects from these sources, 5,227 are also selected by IFS and 2,142 are not. The examples selected only by CAIS are close to orthogonal to the validation gradient: the median absolute cosine similarity between their influence function $\mathcal{I}_{\hat{\theta}}(Z_i)$ and $\nabla_\theta F_{\text{val}}(\hat{\theta})$ is 8.61×10^{-4}. These are the $b_i \approx 0$ examples of Section 3.3, which IFS assigns near-floor inclusion probability, and they correspond to typical OCR content such as book covers, titles, and graffiti text. Such examples are difficult for IFS to select, whereas our curvature-aware design retains them. We further analyze how the subset selected by CAIS under the 40K budget is distributed over the sources of the full training set; see Appendix B.2 for details.

Table 4: Performance comparison between CAIS and IFS under three selection budgets.

Method	#Sample	MMBench-En	MMBench-Cn	MME-P	MME-C	AI2D	POPE-A	POPE-P	SQA-IMG	OCRBench	Rel.
Full	625K	61.81	54.15	1319.64	268.75	54.01	83.40	93.72	69.39	14.70	100.00%
IFS	40K	57.55	50.64	1161.58	271.88	50.45	83.92	83.17	67.78	13.90	94.54%
CAIS (Ours)		61.17	52.98	1211.01	297.32	51.82	84.35	87.45	68.84	14.20	98.38%
IFS	100K	60.21	54.26	1202.04	287.05	53.20	84.11	90.93	68.40	14.20	98.56%
CAIS (Ours)		62.77	53.83	1340.83	231.70	53.81	82.93	92.49	69.21	15.20	98.85%
IFS	200K	61.60	54.47	1339.54	266.07	53.36	84.47	89.91	68.34	14.70	99.47%
CAIS (Ours)		62.98	55.21	1386.76	262.95	52.71	84.43	91.49	66.79	14.60	99.87%

Continued training with selected data In practice, visual instruction tuning often runs for multiple epochs rather than stopping after a single pass over the data. We therefore study the role of our selection method in such multi-epoch training. Table 5 compares full data training for $r \in \{2, 3\}$ epochs against a hybrid schedule: one epoch on the full dataset followed by $r - 1$ epochs on a selected subset (with a budget of 40K examples). While both configurations share the same total number of epochs, each subsequent subset epoch processes approximately 40K examples instead of the full 625K, a reduction of 585K examples (93.6%). Under the same effective batch size, the number of optimizer updates per epoch is reduced by the same proportion. Reported compute savings exclude feature extraction and selection.

Replacing subsequent full data epochs with CAIS subset training substantially reduces training cost while preserving aggregate performance. Over two epochs, Full + CAIS achieves a Rel. of 102.47%, only 0.44 percentage points below Full, while reducing training GPU-hours by 47.04%. Over three epochs, it reaches 103.37%, exceeding Full by 1.38 points while reducing training GPU-hours by 62.88%. These results demonstrate the utility of the selected subset for continued training with substantially less training compute. In a separate 40K timing run, data selection takes 8.53 hours in total. This one-time cost is modest compared with multi-epoch model fine-tuning.

Table 5: Performance of CAIS under continued training.

Epochs	Training Dataset†	MMBench-En	MMBench-Cn	MME-P	MME-C	AI2D	POPE-A	POPE-P	SQA-IMG	OCRBench	Rel.	GPU-h	Savings (%)
1	full data	61.81	54.15	1319.64	268.75	54.01	83.40	93.72	69.39	14.70	100.00%	—	—
2	full data × 2	61.17	56.49	1405.14	339.73	52.67	84.11	93.55	66.54	14.10	102.91%	198.69	47.04%
	full data + CAIS	62.13	55.43	1321.07	353.57	51.86	84.89	88.97	67.78	14.30	102.47%	105.23	
3	full data × 3	61.38	55.43	1372.81	329.46	52.63	84.74	94.18	64.87	14.20	101.99%	297.46	62.88%
	full data + CAIS × 2	61.81	54.89	1388.43	373.66	52.39	83.78	93.13	67.10	13.40	103.37%	110.40	

† $\times r$: trained on the corresponding data for r epochs.

5 Related Work

Influence estimation and gradient-based data selection Influence functions measure how perturbing the weight on a training example changes the fitted parameters and, through the chain rule, a target loss (Cook & Weisberg, 1982; Koh & Liang, 2017). A second line of work turns such scores into selection rules: LESS (Xia et al., 2024) ranks candidates by Adam-aware gradient similarity to a validation set, Influence Distillation (Nikdan et al., 2025) propagates second-order sample weights from a small landmark set, and DiReCT (Xu et al., 2026) constrains per sample gradients using the spectral geometry of validation curvature. Others reduce the cost of the scores rather than their use, via influence preserving proxy models (Chen et al., 2026) or training-free multimodal selection (Wu et al., 2026).

Efficient Hessian computation Influence functions require inverse-Hessian-vector products (IHVPs), and explicitly inverting the Hessian is infeasible for modern networks. Existing methods approximate them with iterative solvers built on exact Hessian-vector products (Pearlmutter, 1994; Agarwal et al., 2017; Schioppa et al., 2022), Kronecker-factored curvature (Martens & Grosse, 2015; George et al., 2018), which scales influence estimation to large language models (Grosse et al., 2023), or random projections of per-example gradients (Park et al., 2023; Choe et al., 2024). These approximations can deviate substantially from leave-one-out retraining in deep networks (Basu et al., 2021; Bae et al., 2022). We adopt the projected route, but compute a quadratic form of the influence under the validation curvature rather than an IHVP.

Optimal subsampling design In statistics, non-uniform subsampling is posed as choosing inclusion probabilities that minimize the asymptotic variance of an estimator computed on the sample. For logistic regression, Wang et al. (2018) derive A- and L-optimal subsampling probabilities from the asymptotic distribution of the subsample estimator, showing that they depend on both the residual and the covariate leverage. Ting & Brochu (2018) extend this beyond specific parametric models: for asymptotically linear estimators the influence function yields optimal Poisson probabilities proportional to the influence magnitude, with Horvitz–Thompson reweighting to remove selection bias. Their optimality is stated for the variance of a first-order correction to the estimator, which under Horvitz–Thompson weighting has zero mean. The resulting designs are therefore driven by gradient magnitude, and leave open how an example's influence interacts with loss curvature.

6 Conclusion

We formulate budget constrained data selection as the design of Poisson inclusion probabilities that minimize the expected validation risk after subset retraining. We combine a linearization of the weighted training-loss gradient with a second-order expansion of validation risk to derive a curvature-aware risk surrogate with a closed-form optimal design. The excess risk of this design relative to the optimal design decays as $o(B^{-1})$, and influence function based sampling is recovered as its first-order special case. With gradient sketching and Gauss Newton curvature estimates, the design scales to modern vision language models; on LLaVA-Vicuna-7B it retains 98.38% of full-data performance using only 6.4% of the candidate set. Several directions remain for future work. Our guarantee concerns exact coefficients, and extending it to the sketched scores used in practice would close the gap between theory and implementation. In addition, the design is computed once at a fixed reference checkpoint; updating the curvature and inclusion probabilities during training may further benefit multi-epoch selection and continued fine-tuning under a fixed compute budget.

AI USE STATEMENT

We used generative AI tools only to improve the clarity and readability of the manuscript. The authors reviewed and revised all AI-assisted edits and take full responsibility for the content of this paper. We did not use generative AI for the research methodology, experiments, data processing, analysis, or interpretation of results.

ETHICS STATEMENT

This work studies data selection for visual instruction tuning using existing datasets and models. We did not collect new human-subject data. The selected data and resulting models may inherit biases and other limitations of the underlying datasets and models.

REPRODUCIBILITY STATEMENT

The experimental setup, selection and retraining procedure, training hyperparameters, and random seeds are documented in Section 4.1 and Appendix B.1. We plan to release the implementation upon publication.

REFERENCES

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *Journal of Machine Learning Research*, 18(116):1–40, 2017.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems*, volume 35, pp. 17953–17967, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Samyadeep Basu, Phillip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL `https://web.stanford.edu/~boyd/cvxbook/`.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Automata, Languages and Programming*, 2002.

Sirui Chen, Yunzhe Qi, Mengting Ai, Yifan Sun, Ruizhong Qiu, Jiaru Zou, and Jingrui He. Influence-preserving proxies for gradient-based data selection in LLM fine-tuning. In *International Conference on Learning Representations*, 2026.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric P. Xing. What is your data worth to GPT? LLM-scale data valuation with influence functions. *arXiv preprint arXiv:2405.13954*, 2024.

R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, New York, 1982.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a Kronecker-factored eigenbasis. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pp. 235–251, 2016.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *International Conference on Learning Representations*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1587–1606, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/TMM-OpenWorld/html/Li_A_Survey_of_State_of_the_Art_Large_Vision_Language_CVPRW_2025_paper.html.

Shijie Lian, Changti Wu, Laurence Tianruo Yang, Hang Yuan, Bin Yu, Lei Zhang, and Kai Chen. Euclid's gift: Enhancing spatial perception and reasoning in vision-language models via geometric surrogate tasks. *arXiv preprint arXiv:2509.24473*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233, 2024c.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024d. doi: 10.1007/s11432-024-4235-6.

Zheng Liu, Honglin Lin, Chonghan Qin, Xiaoyang Wang, Xin Gao, Yu Li, Mengzhang Cai, Yun Zhu, Zhanping Zhong, Qizhi Pei, et al. ChartVerse: Scaling chart reasoning via reliable programmatic synthesis from scratch. *arXiv preprint arXiv:2601.13606*, 2026.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417, 2015.

Mahdi Nikdan, Vincent Cohen-Addad, Dan Alistarh, and Vahab Mirrokni. Efficient data selection at scale via influence distillation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning*, pp. 27074–27113, 2023.

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8179–8186, 2022.

Daniel Ting and Eric Brochu. Optimal subsampling with influence functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/57c0531e13f40b91b3b0f1a30b529a1d-Paper.pdf.

HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018. doi: 10.1080/01621459.2017.1292914.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

Changti Wu, Shijie Lian, Zihao Liu, Lei Zhang, Laurence Tianruo Yang, and Kai Chen. Dyna-SolidGeo: A dynamic benchmark for genuine spatial mathematical reasoning of VLMs in solid geometry. *arXiv preprint arXiv:2510.22340*, 2025.

Changti Wu, Jiahuai Mao, Yuzhuo Miao, Shijie Lian, Bin Yu, Xiaopeng Lin, Cong Huang, Lei Zhang, and Kai Chen. ScalSelect: Scalable training-free multimodal data selection for efficient visual instruction tuning. *arXiv preprint arXiv:2602.11636*, 2026.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Yuanjian Xu, Jianing Hao, Wanbo Zhang, Zhong Li, and Guang Zhang. Towards efficient LLMs annealing with principled sample selection. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.

Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. SVIT: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.

A SUPPORTING RESULTS AND PROOFS

This appendix proves Theorem 1, compares the proposed design with uniform and influence-based sampling, and examines when the assumptions hold. The main line of argument is Appendices A.1–A.3; readers interested only in the proof of the theorem can stop after Appendix A.2.

- Appendix A.1 fixes the setting and states the two assumptions of the theorem.
- Appendix A.2 shows that $\pi_\alpha^\star$ exactly minimizes the surrogate, bounds the gap between the surrogate and the actual risk, and combines the two into the proof of the theorem. The analysis allows a nonstationary warm-up checkpoint θ_0 and inexact retraining.
- Appendix A.3 shows that uniform sampling and scalar influence-based sampling incur an excess risk of order B^{-1}, including Proposition 5 cited in the main text and an explicit example in which both baselines are strictly worse.
- Appendix A.4 derives the assumptions from properties of the losses, verifies them for regularized frozen-feature softmax, and extends the guarantee to random datasets, including a binary cross-entropy example with vanishing validation gradient.
- Appendix A.5 extends the guarantee to the approximate coefficients $\widehat{a}_i$ computed in Section 3.4.

Notation Besides the notation of Sections 2 and 3, we use the following symbols throughout.

Symbol	Meaning
θ_0	reference (warm-up) checkpoint; $\theta_0 = \hat{\theta}$ in Theorem 1
H	reference training Hessian $H_{\mathrm{tr}}(\theta_0)$, with $H \succeq \mu I$
Δ_π, $\Delta_\pi^{\mathrm{lin}}$	actual and linearized training displacement from θ_0
e_π	linearization error $\Delta_\pi - \Delta_\pi^{\mathrm{lin}}$
g_π^{c}, E_π	centered gradient and Hessian perturbations at θ_0
r_0, r_π	optimization residuals before and after subset training
$\overline{H}_\pi$	sampled training Hessian averaged along the retraining segment
$\mathcal{A}_\pi$	stability event $\{\|\overline{H}_\pi - H\|_2 \le \lambda_{\min}(H)/2\}$
$\mathcal{P}(B)$	feasible designs without a probability floor
π^{unif}, π_α^{IF}	uniform design and scalar influence-based design
$\widetilde{a}_i$, $\widehat{\pi}$	effective approximate coefficients and the resulting design

Section	Contents	Page
A.1	Setup and assumptions	13
A.2	Surrogate optimality and proof of Theorem 1	14
A.3	Risk comparisons with uniform and influence-based sampling	21
A.4	Sufficient conditions and examples	25
A.5	Sensitivity to score and optimization errors	39

A.1 SETUP AND ASSUMPTIONS

Training procedure and expectations Unless a result explicitly specializes the setting, θ_0 denotes a fixed warm-up checkpoint and $\hat{\theta}_\pi(W)$ denotes the output of HT-weighted training started from θ_0, using the same optimizer and stopping rule for every design; an empty draw returns θ_0. Theorem 1 is the special case $\theta_0 = \hat{\theta}$ with the exact retraining convention of Section 2.2. In that case both optimization residuals in Eq. A.3 vanish, so Assumption 2 holds automatically. Expectations are conditional on the data, the warm-up, and the score construction, and they average over the sampling indicators and any subsequent optimizer randomness. When retraining is exact, it is a deterministic function of the indicators, and we write $\mathbb{E}_W$.

Designs and the probability floor We consider a sequence of problems indexed by $B \to \infty$, with $B \le n$. Besides the floored class $\mathcal{P}_\alpha(B)$ in Eq. 6, we use the class without a lower bound,

$$\mathcal{P}(B) := \left\{ \pi \in (0,1]^n : \sum_{i=1}^n \pi_i \le B \right\}. \tag{A.1}$$

Throughout the theoretical analysis, the probability floor in Eq. 6 is set to

$$\alpha = cB/n, \qquad 0 < c < 1, \tag{A.2}$$

following the probability regularization of Ting & Brochu (2018), with c fixed independently of B and n. Since uniform sampling assigns probability B/n to every example, the floor guarantees each example at least a fraction c of its uniform probability. It also bounds the HT weight of each selected example by $1/(n\pi_i) \le 1/(cB)$, and $\alpha n = cB < B$ keeps $\mathcal{P}_\alpha(B)$ nonempty for every $0 < B \le n$.

Displacement, residuals, and the stability event Relating the surrogate to the actual risk requires controlling how sampling perturbs the gradient and curvature, and how much optimization error remains after finite training. Write

$$H := H_{\mathrm{tr}}(\theta_0), \qquad \Delta_\pi := \hat{\theta}_\pi(W) - \theta_0$$

for the reference Hessian and the training displacement. The residuals r_0 and r_π measure the departure from stationarity before and after subset training, and $\overline{H}_\pi$ averages the sampled curvature along the segment from θ_0 to $\hat{\theta}_\pi(W)$:

$$r_0 := \nabla_\theta F_{\mathrm{tr}}(\theta_0), \qquad r_\pi := \nabla_\theta F_{\mathrm{tr},\pi}(\hat{\theta}_\pi(W); W), \tag{A.3}$$

$$\overline{H}_\pi := \int_0^1 \nabla^2_\theta F_{\mathrm{tr},\pi}(\theta_0 + t\Delta_\pi; W)\, \mathrm{d}t. \tag{A.4}$$

The segment is an analytical device and need not follow the optimization trajectory. The sampling perturbation at θ_0 is measured by the centered gradient g^{c}_π defined in Eq. A.9. We call an output *stable* when its averaged curvature stays close to H, that is, on the event

$$\mathcal{A}_\pi := \left\{ \|\overline{H}_\pi - H\|_2 \le \tfrac{1}{2}\lambda_{\min}(H) \right\}. \tag{A.5}$$

Assumptions All constants below are independent of B. Theorem 1 imposes both assumptions separately for $\pi = \pi^\star_\alpha$ and $\pi = \pi_{\mathrm{opt}}$; results about a single design impose them under that design's sampling and training laws.

Assumption 1 (Local stability and smoothness)**.** *The training objective is twice continuously differentiable near each output segment, and $H \succeq \mu I$ for a fixed $\mu > 0$. The validation loss is twice continuously differentiable there, its Hessian is L_{val}-Lipschitz in operator norm, and $\|H_{\mathrm{val}}(\theta_0)\|_2$ is bounded. Gradient and curvature perturbations satisfy the fourth-moment bounds*

$$\mathbb{E}\|g^{\mathrm{c}}_\pi\|_2^4 = O(B^{-2}), \qquad \mathbb{E}[\|\overline{H}_\pi - H\|_2^4 \mathbf{1}_{\mathcal{A}_\pi}] = O(B^{-2}). \tag{A.6}$$

Unstable outputs satisfy the tail bounds

$$\mathbb{P}(\mathcal{A}^c_\pi) = O(B^{-2}), \qquad \mathbb{E}[\|\Delta_\pi\|_2^2 \mathbf{1}_{\mathcal{A}^c_\pi}] = O(B^{-2}),\ \mathbb{E}[\|\Delta_\pi\|_2^3 \mathbf{1}_{\mathcal{A}^c_\pi}] \ = O(B^{-3/2}). \tag{A.7}$$

Assumption 2 (Optimization accuracy)**.** *The reference and terminal residuals satisfy*

$$\|r_0\|_2 = O(B^{-1/2}), \qquad (\mathbb{E}\|r_\pi\|_2^4)^{1/4} = O(B^{-1}). \tag{A.8}$$

Interpretation of the conditions The stability event requires the averaged training curvature between θ_0 and the fine-tuned parameters to remain close to the curvature at θ_0. Together with $H \succeq \mu I$, this prevents gradient perturbations from being excessively amplified into parameter changes. The fourth-moment bounds in Eq. A.6 state that sampling perturbations shrink at the $B^{-1/2}$ scale, which is the scale produced by the probability floor (Appendix A.4.3). The tail bounds in Eq. A.7 allow occasional unstable outputs, provided they are rare and their displacements do not dominate the expected risk. Finally, Assumption 2 requires neither stage to be exactly stationary: the warm-up residual may be as large as the sampling perturbation, $O(B^{-1/2})$, while the residual after subset fine-tuning must be $O(B^{-1})$ in the fourth-moment sense.

A.2 Surrogate optimality and proof of Theorem 1

We first identify the surrogate-optimal design, then bound the difference between actual and surrogate risks. These results yield the excess-risk guarantee in Theorem 1.

A.2.1 SURROGATE IDENTITY AND OPTIMAL SAMPLING

For the training procedure specified in Appendix A.1, define its centered gradient perturbation at the fixed checkpoint by

$$g_\pi^{\mathrm{c}} := \nabla_\theta F_{\mathrm{tr},\pi}(\theta_0; W) - \nabla_\theta F_{\mathrm{tr}}(\theta_0) = \frac{1}{n}\sum_i \left(\frac{W_i}{\pi_i} - 1\right) \nabla_\theta \ell(Z_i; \theta_0). \tag{A.9}$$

With invertible $H_{\mathrm{tr}}(\theta_0)$, the local response is

$$\begin{aligned} \mathcal{I}_{\theta_0}(Z_i) &:= -H_{\mathrm{tr}}(\theta_0)^{-1} \nabla_\theta \ell(Z_i; \theta_0), \\ \Delta_\pi^{\mathrm{lin}} &:= -H_{\mathrm{tr}}(\theta_0)^{-1} g_\pi^{\mathrm{c}} = \frac{1}{n}\sum_i \left(\frac{W_i}{\pi_i} - 1\right) \mathcal{I}_{\theta_0}(Z_i). \end{aligned} \tag{A.10}$$

At a stationary reference, this response agrees with the parameter influence in Eq. 3. At a nonstationary checkpoint it is a preconditioned gradient, without an assertion that it is the derivative of a fitted estimator. Define the exact coefficients and the quadratic surrogate at this same checkpoint by

$$a_i := \mathcal{I}_{\theta_0}(Z_i)^\top H_{\mathrm{val}}(\theta_0) \mathcal{I}_{\theta_0}(Z_i), \tag{A.11}$$

$$\widetilde{\mathcal{R}}(\pi) := F_{\mathrm{val}}(\theta_0) + \frac{1}{2n^2}\sum_i \left(\frac{1}{\pi_i} - 1\right) a_i. \tag{A.12}$$

These extend Eq. 12 from $\hat{\theta}$ to θ_0. For positive coefficients and $0 < B < n$, the corresponding rule without a probability floor is

$$\pi_i^\star = \min\{1, \kappa\sqrt{a_i}\}, \qquad \sum_i \pi_i^\star = B. \tag{A.13}$$

This rule minimizes the surrogate over $\mathcal{P}(B)$ after removing the lower-bound constraint. With that constraint, the optimal rule is Eq. 14, with coefficients evaluated at θ_0.

Proposition 1 (Exact surrogate risk and optimal sampling)**.** *For every $\pi \in \mathcal{P}(B)$, $\mathbb{E}[\Delta_\pi^{\mathrm{lin}}] = 0$ and*

$$\begin{aligned} &\mathbb{E}\left[\nabla_\theta F_{\mathrm{val}}(\theta_0)^\top \Delta_\pi^{\mathrm{lin}} + \frac{1}{2}(\Delta_\pi^{\mathrm{lin}})^\top H_{\mathrm{val}}(\theta_0) \Delta_\pi^{\mathrm{lin}}\right] \\ &\qquad = \widetilde{\mathcal{R}}(\pi) - F_{\mathrm{val}}(\theta_0). \end{aligned} \tag{A.14}$$

If every $a_i > 0$ and $0 < B < n$, the probabilities in Eq. A.13 uniquely minimize $\widetilde{\mathcal{R}}$ over $\mathcal{P}(B)$. For positive coefficients, $0 < \alpha < 1$, and $\alpha n \le B \le n$, the floored rule in Eq. 14, with coefficients evaluated at θ_0, uniquely minimizes $\widetilde{\mathcal{R}}$ over $\mathcal{P}_\alpha(B)$. No stationarity of the validation loss is required.

HT weighting removes the first-order sampling effect in expectation under the linearized response. The remaining quadratic term measures sampling variability through validation curvature. The floored class $\mathcal{P}_\alpha(B)$ is defined in Eq. 6.

Proof. For a Bernoulli inclusion indicator with probability $\pi_i > 0$,

$$\mathbb{E}\left(\frac{W_i}{\pi_i} - 1\right) = \pi_i\left(\frac{1}{\pi_i} - 1\right) - (1 - \pi_i) = 0, \tag{A.15}$$

$$\mathbb{E}\left(\frac{W_i}{\pi_i} - 1\right)^2 = \pi_i\left(\frac{1}{\pi_i} - 1\right)^2 + (1 - \pi_i) = \frac{1}{\pi_i} - 1. \tag{A.16}$$

For $i \neq j$, independence gives

$$\mathbb{E}\left[\left(\frac{W_i}{\pi_i} - 1\right)\left(\frac{W_j}{\pi_j} - 1\right)\right] = \mathbb{E}\left(\frac{W_i}{\pi_i} - 1\right)\mathbb{E}\left(\frac{W_j}{\pi_j} - 1\right) = 0. \tag{A.17}$$

Linearity of expectation in Eq. A.10 gives $\mathbb{E}[\Delta_\pi^{\mathrm{lin}}] = 0$. Since the validation gradient is fixed before sampling,

$$\mathbb{E}\big[\nabla_\theta F_{\mathrm{val}}(\theta_0)^\top \Delta_\pi^{\mathrm{lin}}\big] = \nabla_\theta F_{\mathrm{val}}(\theta_0)^\top \mathbb{E}[\Delta_\pi^{\mathrm{lin}}] = 0. \tag{A.18}$$

The quadratic term is a finite double sum. Its expectation is

$$\begin{aligned}
&\mathbb{E}\big[(\Delta_\pi^{\mathrm{lin}})^\top H_{\mathrm{val}}(\theta_0)\Delta_\pi^{\mathrm{lin}}\big] \\
&\quad= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}\left[\left(\frac{W_i}{\pi_i}-1\right)\left(\frac{W_j}{\pi_j}-1\right)\right]\mathcal{I}_{\theta_0}(Z_i)^\top H_{\mathrm{val}}(\theta_0)\mathcal{I}_{\theta_0}(Z_j) \\
&\quad= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{\pi_i}-1\right)a_i. \qquad \text{(A.19)}
\end{aligned}$$

The last line uses the zero cross moments in Eq. A.17, the diagonal moments in Eq. A.16, and the definition of a_i in Eq. A.11. Multiplying this identity by $1/2$, adding Eq. A.18, and using Eq. A.12 proves Eq. A.14.

For the optimization claim, suppose every $a_i > 0$ and $0 < B < n$. The terms of $\widetilde{\mathcal{R}}$ that depend on π are a positive multiple of $\sum_i a_i/\pi_i$. The function $\kappa \mapsto \sum_i \min(1, \kappa\sqrt{a_i})$ is continuous, equals zero at $\kappa = 0$, and equals n for sufficiently large κ. The intermediate value theorem supplies $\kappa > 0$ for which this sum is B, defining $\pi^\star$ as in Eq. A.13. Set $\lambda = \kappa^{-2}$ and $\gamma_i = \max(0, a_i - \lambda)$. If $\pi_i^\star < 1$, then $a_i/(\pi_i^\star)^2 = \lambda$ and $\gamma_i = 0$. If $\pi_i^\star = 1$, then $a_i \geq \lambda$ and $\gamma_i = a_i - \lambda$. Thus, in either case,

$$-\frac{a_i}{(\pi_i^\star)^2} + \lambda + \gamma_i = 0, \qquad \gamma_i \geq 0, \qquad \gamma_i(\pi_i^\star - 1) = 0. \qquad \text{(A.20)}$$

For any $\pi \in \mathcal{P}(B)$, convexity of a_i/π_i gives

$$\begin{aligned}
\sum_i \frac{a_i}{\pi_i} - \sum_i \frac{a_i}{\pi_i^\star} &\geq \sum_i \frac{-a_i}{(\pi_i^\star)^2}(\pi_i - \pi_i^\star) \\
&= -\lambda\left(\sum_i \pi_i - B\right) - \sum_i \gamma_i(\pi_i - \pi_i^\star) \geq 0. \qquad \text{(A.21)}
\end{aligned}$$

The first term is nonnegative because $\sum_i \pi_i \leq B$. Each nonzero γ_i has $\pi_i^\star = 1$, so the corresponding contribution satisfies $-\gamma_i(\pi_i - \pi_i^\star) \geq 0$ because $\pi_i \leq 1$. Finally, the Hessian of $\sum_i a_i/\pi_i$ is $\mathrm{diag}(2a_i/\pi_i^3) \succ 0$ on the positive orthant. Strict convexity on the convex feasible set makes the minimizer unique. □

The probability floor For $\alpha n \leq B \leq n$, the set $\mathcal{P}_\alpha(B)$ is nonempty and compact. The surrogate is continuous on this set since $\pi_i \geq \alpha > 0$. It therefore attains its minimum at some $\pi_\alpha^\star$, including when coefficients vanish. For positive coefficients and $\alpha n < B < n$, continuity of $\kappa \mapsto \sum_i \max\{\alpha, \min(1, \kappa\sqrt{a_i})\}$ supplies a value with sum B. The resulting probabilities satisfy the KKT conditions with

$$\lambda = \kappa^{-2}, \qquad \nu_i = \max\left(0, \lambda - \frac{a_i}{\alpha^2}\right), \qquad \gamma_i = \max(0, a_i - \lambda), \qquad \text{(A.22)}$$

where ν_i is the multiplier for the lower bound. If $\pi_{\alpha,i}^\star = \alpha$, clipping gives $a_i \leq \lambda\alpha^2$, so $\nu_i = \lambda - a_i/\alpha^2$ and $\gamma_i = 0$. If $\alpha < \pi_{\alpha,i}^\star < 1$, then $a_i/(\pi_{\alpha,i}^\star)^2 = \lambda$ and both multipliers are zero. If $\pi_{\alpha,i}^\star = 1$, then $a_i \geq \lambda$, so $\gamma_i = a_i - \lambda$ and $\nu_i = 0$. In all three cases,

$$-\frac{a_i}{(\pi_{\alpha,i}^\star)^2} + \lambda - \nu_i + \gamma_i = 0, \qquad \nu_i(\pi_{\alpha,i}^\star - \alpha) = 0, \qquad \gamma_i(\pi_{\alpha,i}^\star - 1) = 0.$$

To verify sufficiency, the convexity inequality in Eq. A.21 now has the right-hand side

$$-\lambda\left(\sum_i \pi_i - B\right) + \sum_i \nu_i(\pi_i - \pi_{\alpha,i}^\star) - \sum_i \gamma_i(\pi_i - \pi_{\alpha,i}^\star).$$

The budget term is nonnegative. Where $\nu_i > 0$, $\pi_{\alpha,i}^\star = \alpha \leq \pi_i$; where $\gamma_i > 0$, $\pi_{\alpha,i}^\star = 1 \geq \pi_i$. All remaining terms are nonnegative as well. This proves the stated clipped rule and, by strict convexity, its uniqueness for positive coefficients. At $B = \alpha n$, the only feasible vector has all entries α; at $B = n$, positive coefficients uniquely favor all entries equal to one.

A.2.2 RISK DECOMPOSITION FOR A FIXED DESIGN

Fix a design $\pi \in \mathcal{P}(B)$ before drawing W. With the actual displacement Δ_π defined in Appendix A.1, write its linearization error as

$$e_\pi := \Delta_\pi - \Delta_\pi^{\mathrm{lin}}. \tag{A.23}$$

Proposition 2 (Risk approximation for a fixed design)**.** *Suppose F_{val} is twice continuously differentiable on a neighborhood of the segments joining θ_0 to all supported $\hat{\theta}_\pi(W)$ under this fixed design, and its Hessian is L_{val}-Lipschitz in operator norm along these segments. Assume the following moments are finite, and define*

$$\begin{aligned}\varepsilon(\pi) := \|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2 \|\mathbb{E}[e_\pi]\|_2 \\ + \|H_{\mathrm{val}}(\theta_0)\|_2 \sqrt{\mathbb{E}\|\Delta_\pi^{\mathrm{lin}}\|_2^2 \, \mathbb{E}\|e_\pi\|_2^2} \\ + \frac{1}{2}\|H_{\mathrm{val}}(\theta_0)\|_2 \mathbb{E}\|e_\pi\|_2^2 + \frac{L_{\mathrm{val}}}{6}\mathbb{E}\|\Delta_\pi\|_2^3.\end{aligned} \tag{A.24}$$

Then this design satisfies

$$|\mathcal{R}(\pi) - \widetilde{\mathcal{R}}(\pi)| \le \varepsilon(\pi). \tag{A.25}$$

The first term retains retraining bias: HT centers $\Delta_\pi^{\mathrm{lin}}$, but need not center e_π. Even when $e_\pi = 0$, the validation Taylor remainder may remain. Proposition 3 below controls the error after accounting for the common optimization drift.

Proof. Condition on the data and construction of the fixed design π. Consider one training output under its sampling and training law. Write $d := \hat{\theta}_\pi(W) - \theta_0$ within this proof. For $t \in [0,1]$, let $\phi(t) = F_{\mathrm{val}}(\theta_0 + td)$. The chain rule gives $\phi'(t) = \nabla_\theta F_{\mathrm{val}}(\theta_0 + td)^\top d$ and $\phi''(t) = d^\top H_{\mathrm{val}}(\theta_0 + td)d$. Integrating $\phi'(t) = \phi'(0) + \int_0^t \phi''(s)\,\mathrm{d}s$ over $t \in [0,1]$ and reversing the order of integration yields

$$\phi(1) = \phi(0) + \phi'(0) + \int_0^1 (1-t)\phi''(t)\,\mathrm{d}t.$$

The derivatives are continuous on this compact interval, which justifies these integrations. We split $\phi''(t) = \phi''(0) + [\phi''(t) - \phi''(0)]$ in the integral. The identity $\int_0^1 (1-t)\,\mathrm{d}t = 1/2$ then gives the exact expansion

$$\begin{aligned}F_{\mathrm{val}}(\theta_0 + d) = F_{\mathrm{val}}(\theta_0) + \nabla_\theta F_{\mathrm{val}}(\theta_0)^\top d + \frac{1}{2} d^\top H_{\mathrm{val}}(\theta_0) d + \mathrm{rem}_\pi, \\ \mathrm{rem}_\pi := \int_0^1 (1-t) d^\top [H_{\mathrm{val}}(\theta_0 + td) - H_{\mathrm{val}}(\theta_0)] d \,\mathrm{d}t.\end{aligned} \tag{A.26}$$

The Hessian Lipschitz condition bounds the absolute integrand by $L_{\mathrm{val}} t(1-t)\|d\|_2^3$. Since $\int_0^1 t(1-t)\,\mathrm{d}t = 1/2 - 1/3 = 1/6$,

$$|\,\mathrm{rem}_\pi\,| \le \frac{L_{\mathrm{val}}}{6}\|d\|_2^3. \tag{A.27}$$

Equation A.23 gives $d = \Delta_\pi^{\mathrm{lin}} + e_\pi$. Symmetry of the validation Hessian gives

$$\begin{aligned}d^\top H_{\mathrm{val}}(\theta_0) d = (\Delta_\pi^{\mathrm{lin}})^\top H_{\mathrm{val}}(\theta_0) \Delta_\pi^{\mathrm{lin}} \\ + 2(\Delta_\pi^{\mathrm{lin}})^\top H_{\mathrm{val}}(\theta_0) e_\pi + e_\pi^\top H_{\mathrm{val}}(\theta_0) e_\pi.\end{aligned}$$

Taking expectations in Eq. A.26, and applying Proposition 1 to the linearized terms, yields

$$\begin{aligned}\mathcal{R}(\pi) - \widetilde{\mathcal{R}}(\pi) = \nabla_\theta F_{\mathrm{val}}(\theta_0)^\top \mathbb{E}[e_\pi] \\ + \mathbb{E}\big[(\Delta_\pi^{\mathrm{lin}})^\top H_{\mathrm{val}}(\theta_0) e_\pi\big] \\ + \frac{1}{2}\mathbb{E}\big[e_\pi^\top H_{\mathrm{val}}(\theta_0) e_\pi\big] + \mathbb{E}[\mathrm{rem}_\pi].\end{aligned} \tag{A.28}$$

The assumed moments make all terms integrable. The first term is bounded in absolute value by $\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2 \|\mathbb{E}[e_\pi]\|_2$. For the cross term, the operator-norm inequality followed by Cauchy–Schwarz gives

$$\begin{aligned}
&\left|\mathbb{E}\left[(\Delta_\pi^{\mathrm{lin}})^\top H_{\mathrm{val}}(\theta_0) e_\pi\right]\right| \\
&\quad \le \|H_{\mathrm{val}}(\theta_0)\|_2 \mathbb{E}[\|\Delta_\pi^{\mathrm{lin}}\|_2 \|e_\pi\|_2] \\
&\quad \le \|H_{\mathrm{val}}(\theta_0)\|_2 \sqrt{\mathbb{E}\|\Delta_\pi^{\mathrm{lin}}\|_2^2 \, \mathbb{E}\|e_\pi\|_2^2}.
\end{aligned} \tag{A.29}$$

The remaining quadratic term has absolute value at most $\|H_{\mathrm{val}}(\theta_0)\|_2 \mathbb{E}\|e_\pi\|_2^2/2$. Equation A.27 also gives $|\mathbb{E}[\mathrm{rem}_\pi]| \le L_{\mathrm{val}} \mathbb{E}\|\hat{\theta}_\pi(W) - \theta_0\|_2^3/6$. Applying the triangle inequality to Eq. A.28 and adding these four bounds proves Eq. A.25. □

Interpreting the error bound One moment in Eq. A.24 is available directly from the local response vectors:

$$\mathbb{E}\|\Delta_\pi^{\mathrm{lin}}\|_2^2 = \frac{1}{n^2} \sum_i \left(\frac{1}{\pi_i} - 1\right) \|\mathcal{I}_{\theta_0}(Z_i)\|_2^2. \tag{A.30}$$

This follows by expanding the squared norm as a double sum and applying Eqs. A.16 and A.17. This identity evaluates the moment under the fixed design π; it does not require a bound over other designs. The remaining moments require control of the training output; training stationarity alone does not supply it. In particular, $\mathbb{E}[\Delta_\pi^{\mathrm{lin}}] = 0$ leaves $\nabla_\theta F_{\mathrm{val}}(\theta_0)^\top \mathbb{E}[e_\pi]$ in Eq. A.28. If $\Delta_\pi = \Delta_\pi^{\mathrm{lin}}$ and the validation loss is quadratic on the relevant segments, all four error terms vanish. For general retraining, Proposition 2 is a conditional guarantee whose usefulness depends on these moments under the fixed design.

A.2.3 Risk approximation with optimization residuals

Use the residuals and event defined in Appendix A.1. The reference residual produces the common displacement and quadratic risk contribution

$$d_0 := -H^{-1} r_0, \qquad C_B := \nabla_\theta F_{\mathrm{val}}(\theta_0)^\top d_0 + \tfrac{1}{2} d_0^\top H_{\mathrm{val}}(\theta_0) d_0. \tag{A.31}$$

Both are fixed before sampling and do not depend on the design.

Proposition 3 (Corrected risk accuracy for a fixed design)**.** *Fix a sequence of designs $\pi \in \mathcal{P}(B)$, with $B \to \infty$ and $B \le n$. Suppose Assumption 1 holds under this sequence's sampling and training laws. Assume*

$$\|r_0\|_2 = O(B^{-1/2}), \qquad (\mathbb{E}\|r_\pi\|_2^4)^{1/4} = O(\rho_B), \qquad 0 \le \rho_B \le C B^{-1/2}, \tag{A.32}$$

with all constants independent of B. Then

$$\begin{aligned}
|\mathcal{R}(\pi) - \widetilde{\mathcal{R}}(\pi) - C_B| = O\Bigg(&\frac{\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2}{B} + B^{-3/2} \\
&+ \left(\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2 + B^{-1/2}\right)\rho_B + \rho_B^2 \Bigg).
\end{aligned} \tag{A.33}$$

The result concerns this fixed design sequence. It imposes no score-accuracy condition and requires no independence between optimization residuals and sampling perturbations.

Proof. Fix the design sequence in the proposition. For this proof, let $q := B^{-1/2}$ and choose a constant $\mu > 0$ such that $H \succeq \mu I$. Define the remaining error $\beta_\pi := e_\pi - d_0$. These definitions give

$$\Delta_\pi = d_0 + \Delta_\pi^{\mathrm{lin}} + \beta_\pi, \qquad \|d_0\|_2 \le \mu^{-1} \|r_0\|_2 = O(q). \tag{A.34}$$

Displacement and error on the stable event Integrating the sampled training gradient along the output segment gives the exact identity

$$r_\pi = r_0 + g_\pi^{\mathrm{c}} + \overline{H}_\pi \Delta_\pi. \tag{A.35}$$

Since $H\Delta_\pi^{\mathrm{lin}} = -g_\pi^{\mathrm{c}}$ and $Hd_0 = -r_0$, subtraction yields

$$H\beta_\pi = r_\pi + (H - \overline{H}_\pi)\Delta_\pi. \tag{A.36}$$

On $\mathcal{A}_\pi$, the event's spectral bound gives $\lambda_{\min}(\overline{H}_\pi) \geq \lambda_{\min}(H) - \|\overline{H}_\pi - H\|_2 \geq \mu/2$; thus Eq. A.35 implies

$$\|\Delta_\pi\|_2 \leq \frac{2}{\mu}(\|r_0\|_2 + \|g_\pi^{\mathrm{c}}\|_2 + \|r_\pi\|_2).$$

For nonnegative x, y, z, convexity of the fourth power gives $(x+y+z)^4 \leq 27(x^4+y^4+z^4)$. Consequently,

$$\begin{aligned}\mathbb{E}[\|\Delta_\pi\|_2^4 \mathbf{1}_{\mathcal{A}_\pi}] &\leq 27(2/\mu)^4(\|r_0\|_2^4 + \mathbb{E}\|g_\pi^{\mathrm{c}}\|_2^4 + \mathbb{E}\|r_\pi\|_2^4) \\ &= O(q^4) + O(q^4) + O(\rho_B^4) = O(q^4),\end{aligned}$$

where $\rho_B = O(q)$ controls the last term. Squaring Eq. A.36, using $(x+y)^2 \leq 2x^2 + 2y^2$, and applying Cauchy–Schwarz to the curvature–displacement product gives

$$\begin{aligned}\mathbb{E}[\|\beta_\pi\|_2^2 \mathbf{1}_{\mathcal{A}_\pi}] &\leq \frac{2}{\mu^2}(\mathbb{E}\|r_\pi\|_2^4)^{1/2} \\ &\quad + \frac{2}{\mu^2}\big(\mathbb{E}[\|\overline{H}_\pi - H\|_2^4 \mathbf{1}_{\mathcal{A}_\pi}]\big)^{1/2}\big(\mathbb{E}[\|\Delta_\pi\|_2^4 \mathbf{1}_{\mathcal{A}_\pi}]\big)^{1/2} \\ &= O(\rho_B^2) + O(q^2)O(q^2) = O(\rho_B^2 + q^4).\end{aligned}$$

This product bound allows arbitrary dependence among the residual, sampled gradient, and averaged Hessian.

Exceptional outputs and global moments On $\mathcal{A}_\pi^c$, we use the algebraic identity $\beta_\pi = \Delta_\pi - d_0 + H^{-1}g_\pi^{\mathrm{c}}$, without inverting $\overline{H}_\pi$. The squared norm of a sum of three vectors is at most three times the sum of their squared norms, so

$$\begin{aligned}\mathbb{E}[\|\beta_\pi\|_2^2 \mathbf{1}_{\mathcal{A}_\pi^{\mathrm{c}}}] &\leq 3\mathbb{E}[\|\Delta_\pi\|_2^2 \mathbf{1}_{\mathcal{A}_\pi^{\mathrm{c}}}] + 3\|d_0\|_2^2 \mathbb{P}(\mathcal{A}_\pi^c) \\ &\quad + \frac{3}{\mu^2}(\mathbb{E}\|g_\pi^{\mathrm{c}}\|_2^4)^{1/2}\mathbb{P}(\mathcal{A}_\pi^c)^{1/2} \\ &= O(q^4) + O(q^2)O(q^4) + O(q^2)O(q^2) = O(q^4).\end{aligned}$$

The last product follows from Cauchy–Schwarz with the event indicator. Adding the stable and exceptional contributions gives

$$\mathbb{E}\|\beta_\pi\|_2^2 = O(\rho_B^2 + q^4), \qquad (\mathbb{E}\|\beta_\pi\|_2^2)^{1/2} = O(\rho_B + q^2). \tag{A.37}$$

The remaining moments follow from the centered-gradient condition, the stable displacement bound, and the assumed exceptional third moment:

$$\begin{aligned}\mathbb{E}\|\Delta_\pi^{\mathrm{lin}}\|_2^2 &\leq \mu^{-2}(\mathbb{E}\|g_\pi^{\mathrm{c}}\|_2^4)^{1/2} = O(q^2), \\ \mathbb{E}\|\Delta_\pi\|_2^3 &\leq \big(\mathbb{E}[\|\Delta_\pi\|_2^4 \mathbf{1}_{\mathcal{A}_\pi}]\big)^{3/4} + \mathbb{E}[\|\Delta_\pi\|_2^3 \mathbf{1}_{\mathcal{A}_\pi^{\mathrm{c}}}] \\ &= O(q^4)^{3/4} + O(q^3) = O(q^3).\end{aligned}$$

The stable-event inequality uses Hölder with the event indicator; its probability factor is at most one.

Risk expansion around the common displacement For compactness in this paragraph write $v := \nabla_\theta F_{\mathrm{val}}(\theta_0)$ and $V := H_{\mathrm{val}}(\theta_0)$. Independence and centering of the Bernoulli coefficients give

$$\mathbb{E}\Delta_\pi^{\mathrm{lin}} = 0, \qquad \mathbb{E}[(\Delta_\pi^{\mathrm{lin}})^\top V \Delta_\pi^{\mathrm{lin}}] = \frac{1}{n^2}\sum_i \left(\frac{1}{\pi_i} - 1\right) a_i.$$

Indeed, terms with distinct indices have expectation zero, and a diagonal coefficient has second moment $1/\pi_i - 1$, as in Eqs. A.15–A.17. Neither equality uses $r_0 = 0$. The design-independent contribution is C_B as defined above. Expand the validation loss using Eq. A.26, with $\Delta_\pi = d_0 + \Delta_\pi^{\mathrm{lin}} + \beta_\pi$. The terms linear in $\Delta_\pi^{\mathrm{lin}}$ and its cross term with the deterministic d_0 have mean zero. The quadratic $\Delta_\pi^{\mathrm{lin}}$ term gives the surrogate increment. The remaining terms are exactly

$$\begin{aligned}\mathcal{R}(\pi) - \widetilde{\mathcal{R}}(\pi) - C_B = (v + Vd_0)^\top \mathbb{E}\beta_\pi + \mathbb{E}[(\Delta_\pi^{\mathrm{lin}})^\top V \beta_\pi] \\ + \tfrac{1}{2}\mathbb{E}[\beta_\pi^\top V \beta_\pi] + \mathbb{E}\,\mathrm{rem}_\pi\,. \end{aligned} \tag{A.38}$$

All terms are integrable by the preceding moments. The Taylor remainder is bounded by $L_{\mathrm{val}}\|\Delta_\pi\|_2^3/6$. The triangle inequality and Cauchy–Schwarz yield

$$\begin{aligned}|\mathcal{R}(\pi) - \widetilde{\mathcal{R}}(\pi) - C_B| &\le (\|v\|_2 + \|V\|_2\|d_0\|_2)(\mathbb{E}\|\beta_\pi\|_2^2)^{1/2} \\ &\quad + \|V\|_2(\mathbb{E}\|\Delta_\pi^{\mathrm{lin}}\|_2^2)^{1/2}(\mathbb{E}\|\beta_\pi\|_2^2)^{1/2} \\ &\quad + \tfrac{1}{2}\|V\|_2\mathbb{E}\|\beta_\pi\|_2^2 + \tfrac{L_{\mathrm{val}}}{6}\mathbb{E}\|\Delta_\pi\|_2^3 \\ &= (\|v\|_2 + O(q))O(\rho_B + q^2) + O(q)O(\rho_B + q^2) \\ &\quad + O(\rho_B^2 + q^4) + O(q^3) \\ &= O(\|v\|_2\rho_B + \|v\|_2 q^2 + q\rho_B + q^3 + \rho_B^2 + q^4).\end{aligned}$$

Since $q \le 1$ for sufficiently large B, $q^4 \le q^3$. Substituting $q = B^{-1/2}$ proves Eq. A.33. □

Why the common shift matters Since $\|d_0\|_2 = O(B^{-1/2})$ and the validation Hessian is bounded,

$$|C_B| \le \|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2\|d_0\|_2 + \tfrac{1}{2}\|H_{\mathrm{val}}(\theta_0)\|_2\|d_0\|_2^2 = \|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2 O(B^{-1/2}) + O(B^{-1}).$$

The uncorrected absolute risk difference can therefore be larger than the bound in Eq. A.33. Risk comparisons can eliminate C_B because it is common to both designs; HT centering alone does not remove it from absolute risk.

A.2.4 Exact stationary reference

The theorem and the comparisons in Appendix A.3 use the special case $\theta_0 = \hat{\theta}$, where $\hat{\theta}$ is the full-data fit, so $\nabla_\theta F_{\mathrm{tr}}(\hat{\theta}) = 0$. Retraining returns a finite minimizer of the sampled objective, chosen by a deterministic rule when minimizers are not unique, and $\hat{\theta}_\pi(W) = \hat{\theta}$ on the empty draw. Thus $r_0 = 0$, $r_\pi = 0$, and $C_B = 0$. We call this the *exact-stationarity setting*; it is a special case and does not assert that a short warm-up attains stationarity.

Corollary A.1 (Risk accuracy at an exact stationary reference)**.** *Fix a design sequence $\pi \in \mathcal{P}(B)$ satisfying Assumption 1 in the exact-stationarity setting. Then*

$$|\mathcal{R}(\pi) - \widetilde{\mathcal{R}}(\pi)| = O\left(\frac{\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2}{B} + B^{-3/2}\right). \tag{A.39}$$

Proof. The centered gradient equals the sampled gradient because $r_0 = 0$. Proposition 3 applies with $\rho_B = 0$. Its common shift vanishes, and its terms containing ρ_B are zero, giving the displayed bound. Its proof also gives, since $\beta_\pi = e_\pi$,

$$\mathbb{E}\|e_\pi\|_2^2 = O(B^{-2}), \quad \mathbb{E}\|\Delta_\pi^{\mathrm{lin}}\|_2^2 = O(B^{-1}), \quad \mathbb{E}\|\Delta_\pi\|_2^3 = O(B^{-3/2}).$$

In Eq. A.24, the mean error is at most $(\mathbb{E}\|e_\pi\|_2^2)^{1/2} = O(B^{-1})$; the remaining terms have orders $O(B^{-1/2})O(B^{-1}) = O(B^{-3/2})$, $O(B^{-2})$, and $O(B^{-3/2})$. Their sum also bounds $\varepsilon(\pi)$ at the stated rate. □

A.2.5 Proof of Theorem 1

Proof. We first establish the comparison at a general fixed checkpoint θ_0 under Assumptions 1–2, then specialize to the stationary setting of the theorem. For each design $\pi = \pi_\alpha^\star$ and $\pi = \pi_{\mathrm{opt}}$, apply Proposition 3 with $\rho_B = B^{-1}$. The three terminal-residual contributions are

$$\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2 O(B^{-1}), \qquad O(B^{-1/2})O(B^{-1}) = O(B^{-3/2}), \qquad O(B^{-1})^2 = O(B^{-2}).$$

Since $B^{-2} \le B^{-3/2}$ for $B \ge 1$, each corrected risk difference is bounded by a constant times $\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2/B + B^{-3/2}$. The constants may differ at the two designs, but are independent of B. The exact decomposition is

$$\begin{aligned}\mathcal{R}(\pi_\alpha^\star) - \mathcal{R}(\pi_{\mathrm{opt}}) = [\mathcal{R}(\pi_\alpha^\star) - \widetilde{\mathcal{R}}(\pi_\alpha^\star) - C_B] \\ + [\widetilde{\mathcal{R}}(\pi_\alpha^\star) - \widetilde{\mathcal{R}}(\pi_{\mathrm{opt}})] \\ - [\mathcal{R}(\pi_{\mathrm{opt}}) - \widetilde{\mathcal{R}}(\pi_{\mathrm{opt}}) - C_B].\end{aligned}$$

Proposition 1 makes the middle bracket nonpositive. Bounding the other two brackets in absolute value and adding their rates gives

$$\begin{aligned}\mathcal{R}(\pi_\alpha^\star) - \mathcal{R}(\pi_{\mathrm{opt}}) &\le O\left(\frac{\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2}{B} + B^{-3/2}\right) + O\left(\frac{\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2}{B} + B^{-3/2}\right) \\ &= O\left(\frac{\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2}{B} + B^{-3/2}\right).\end{aligned}$$

The lower bound follows because $\pi_\alpha^\star$ is feasible and π_{opt} minimizes $\mathcal{R}$ for the same training procedure. For Theorem 1, set $\theta_0 = \hat{\theta}$. Then $r_0 = 0$, so $C_B = 0$; the terminal-residual contributions are already controlled by Assumption 2. The comparison above yields, for some $C > 0$ independent of B and all sufficiently large B,

$$0 \le \mathcal{R}(\pi_\alpha^\star) - \mathcal{R}(\pi_{\mathrm{opt}}) \le C\left[\frac{\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2}{B} + B^{-3/2}\right]. \tag{A.40}$$

This bound does not require the validation gradient to vanish. If $\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2 = O(r_B)$, the gradient term has order $O(r_B)B^{-1} = O(r_B/B)$. Adding the $O(B^{-3/2})$ remainder proves the general bound. When $r_B = o(1)$, the gradient norm is $O(r_B) = o(1)$, so multiplication by B gives

$$0 \le B[\mathcal{R}(\pi_\alpha^\star) - \mathcal{R}(\pi_{\mathrm{opt}})] \le C[\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2 + B^{-1/2}] \longrightarrow 0,$$

proving the $o(B^{-1})$ conclusion. When $r_B = O(B^{-1/2})$, the gradient norm is $O(r_B) = O(B^{-1/2})$, and the two contributions are $O(B^{-1/2})B^{-1} = O(B^{-3/2})$ and $O(B^{-3/2})$. Their sum proves the $O(B^{-3/2})$ conclusion. Dividing by a surrogate increment $\Theta(B^{-1})$, when such a lower bound holds, gives $O(B^{-3/2})/\Theta(B^{-1}) = O(B^{-1/2})$. □

A.3 Risk comparisons with uniform and influence-based sampling

We compare designs that share the expected budget, the probability floor, and the HT-weighted retraining rule. Throughout this subsection, we work in the exact-stationarity setting of Appendix A.2.4 and assume $a_i > 0$ for all i. Each comparison proceeds in two steps: it computes the exact surrogate gap between an alternative design and $\pi_\alpha^\star$, and then transfers this gap to the actual retraining risk using Corollary A.1. The general result allows either probability bound to be active.

A comparison for any stable design

Proposition 4 (Risk difference from the curvature-optimal design)**.** *Fix a sequence of designs $\pi \in \mathcal{P}_\alpha(B)$. Suppose Assumption 1 holds separately for π and $\pi_\alpha^\star$, with constants independent of B. Then*

$$\begin{aligned}\mathcal{R}(\pi) - \mathcal{R}(\pi_\alpha^\star) = \frac{1}{2n^2}\sum_i a_i\left(\frac{1}{\pi_i} - \frac{1}{\pi_{\alpha,i}^\star}\right) \\ + O\left[\frac{\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2}{B} + B^{-3/2}\right].\end{aligned} \tag{A.41}$$

The first term on the right is the nonnegative surrogate risk difference $\widetilde{\mathcal{R}}(\pi) - \widetilde{\mathcal{R}}(\pi_\alpha^\star)$, and vanishes if and only if $\pi = \pi_\alpha^\star$. If $\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2 = o(1)$ and

$$\liminf_{B\to\infty} B[\widetilde{\mathcal{R}}(\pi) - \widetilde{\mathcal{R}}(\pi_\alpha^\star)] > 0, \tag{A.42}$$

then $\mathcal{R}(\pi_\alpha^\star) < \mathcal{R}(\pi)$ for all sufficiently large B.

Proof. By Eq. 12, terms independent of the design cancel, leaving

$$\widetilde{\mathcal{R}}(\pi) - \widetilde{\mathcal{R}}(\pi^\star_\alpha) = \frac{1}{2n^2} \sum_i a_i \left(\frac{1}{\pi_i} - \frac{1}{\pi^\star_{\alpha,i}} \right).$$

For $a_i > 0$, the second derivative of a_i/π_i is $2a_i/\pi_i^3 > 0$. Thus the objective is strictly convex on the convex feasible set, and its minimizer $\pi^\star_\alpha$ is unique. This proves nonnegativity and the equality condition. The exact decomposition

$$\begin{aligned}\mathcal{R}(\pi) - \mathcal{R}(\pi^\star_\alpha) = [\widetilde{\mathcal{R}}(\pi) - \widetilde{\mathcal{R}}(\pi^\star_\alpha)] \\ + [\mathcal{R}(\pi) - \widetilde{\mathcal{R}}(\pi)] - [\mathcal{R}(\pi^\star_\alpha) - \widetilde{\mathcal{R}}(\pi^\star_\alpha)]\end{aligned}$$

and Corollary A.1, applied separately to the two designs, give Eq. A.41. For gradient norm $o(1)$, multiplying the remainder by B gives $O[\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2 + B^{-1/2}] = o(1)$. Condition A.42 supplies a fixed $d > 0$ such that the scaled surrogate gap is at least d for all sufficiently large B. The scaled approximation error is eventually smaller than $d/2$ in absolute value, making the actual risk difference at least $d/(2B) > 0$. □

If the conditions of Theorem 1 also hold for the actual-risk optimizer and $r_B = o(1)$, the decomposition $\mathcal{R}(\pi) - \mathcal{R}(\pi_{\mathrm{opt}}) = [\mathcal{R}(\pi) - \mathcal{R}(\pi^\star_\alpha)] + [\mathcal{R}(\pi^\star_\alpha) - \mathcal{R}(\pi_{\mathrm{opt}})]$ gives

$$\mathcal{R}(\pi) - \mathcal{R}(\pi_{\mathrm{opt}}) = \widetilde{\mathcal{R}}(\pi) - \widetilde{\mathcal{R}}(\pi^\star_\alpha) + o(B^{-1}). \tag{A.43}$$

Thus a surrogate gap $d/B + o(B^{-1})$ with $d > 0$ gives the same leading excess risk relative to the actual-risk optimizer. Distinct designs alone do not guarantee such a gap: their difference may shrink with B. The constants above concern the specified design sequences; no uniform approximation over all feasible designs is assumed.

Uniform sampling

Proposition 5 (Excess validation risk of uniform sampling)**.** *Under the conditions of Theorem 1, with $r_B = o(1)$, suppose Assumption 1 also holds for $\pi_i^{\mathrm{unif}} = B/n$ under the same retraining rule. Assume $\alpha < \pi^\star_{\alpha,i} < 1$ for every i and all sufficiently large B, and $\mathrm{Var}_n(\sqrt{a_i}) := n^{-1}\sum_i a_i - (n^{-1}\sum_i \sqrt{a_i})^2 \to \sigma_a^2 \in (0,\infty)$. Then*

$$\mathcal{R}(\pi^{\mathrm{unif}}) - \mathcal{R}(\pi_{\mathrm{opt}}) = \frac{\sigma_a^2}{2B} + o(B^{-1}) = \Theta(B^{-1}). \tag{A.44}$$

Proof. Since neither probability bound is active, Eq. 14 and the budget constraint give $\pi^\star_{\alpha,i} = B\sqrt{a_i}/\sum_j \sqrt{a_j}$. Substitution into the design-dependent terms yields

$$\sum_i \frac{a_i}{\pi^\star_{\alpha,i}} = \frac{(\sum_i \sqrt{a_i})^2}{B}, \qquad \sum_i \frac{a_i}{\pi_i^{\mathrm{unif}}} = \frac{n}{B}\sum_i a_i.$$

The remaining terms in Eq. 12 are common to both designs, so

$$\begin{aligned}\widetilde{\mathcal{R}}(\pi^{\mathrm{unif}}) - \widetilde{\mathcal{R}}(\pi^\star_\alpha) &= \frac{n\sum_i a_i - (\sum_i \sqrt{a_i})^2}{2n^2 B} \\ &= \frac{\mathrm{Var}_n(\sqrt{a_i})}{2B}.\end{aligned}$$

Applying Proposition 4 to $\pi = \pi^{\mathrm{unif}}$ gives

$$\mathcal{R}(\pi^{\mathrm{unif}}) - \mathcal{R}(\pi^\star_\alpha) = \frac{\mathrm{Var}_n(\sqrt{a_i})}{2B} + O\left[\frac{\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2}{B} + B^{-3/2} \right]. \tag{A.45}$$

Using the exact decomposition

$$\begin{aligned}\mathcal{R}(\pi^{\mathrm{unif}}) - \mathcal{R}(\pi_{\mathrm{opt}}) = [\mathcal{R}(\pi^{\mathrm{unif}}) - \mathcal{R}(\pi^\star_\alpha)] \\ + [\mathcal{R}(\pi^\star_\alpha) - \mathcal{R}(\pi_{\mathrm{opt}})]\end{aligned}$$

and Eq. A.40 yields the finite-budget comparison

$$\mathcal{R}(\pi^{\mathrm{unif}}) - \mathcal{R}(\pi_{\mathrm{opt}}) = \frac{\mathrm{Var}_n(\sqrt{a_i})}{2B} + O\left[\frac{\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2}{B} + B^{-3/2}\right]. \tag{A.46}$$

For gradient norm $o(1)$, the remainder has order $o(1)B^{-1} + O(B^{-3/2}) = o(B^{-1})$. Also, $\mathrm{Var}_n(\sqrt{a_i}) = \sigma_a^2 + o(1)$ implies $\mathrm{Var}_n(\sqrt{a_i})/(2B) = \sigma_a^2/(2B) + o(B^{-1})$. Thus $B[\mathcal{R}(\pi^{\mathrm{unif}}) - \mathcal{R}(\pi_{\mathrm{opt}})] \to \sigma_a^2/2 > 0$, proving both assertions in Eq. A.44. If the gradient norm is $O(B^{-1/2})$, the remainder in Eq. A.46 is $O(B^{-3/2})$; replacing the empirical variance by its limit preserves this rate only if their difference is $O(B^{-1/2})$. □

Scalar influence-based sampling Here IF denotes the scalar validation-loss influence design in Section 3.3, whose coefficients are $b_i = \mathcal{I}(Z_i)^2$. It does not denote every sampling method based on parameter influence vectors.

Corollary A.2 (Comparison with scalar influence-based sampling)**.** *Suppose $b_i > 0$ for all i, and the assumptions of Proposition 4 hold for $\pi = \pi_\alpha^{\mathrm{IF}}$. The actual risk difference is given by Eq. A.41, with a nonnegative surrogate difference that is zero exactly when $\pi_\alpha^{\mathrm{IF}} = \pi_\alpha^\star$. If both designs have strictly interior probabilities, then*

$$\begin{aligned}&\widetilde{\mathcal{R}}(\pi_\alpha^{\mathrm{IF}}) - \widetilde{\mathcal{R}}(\pi_\alpha^\star)\\ &\quad= \frac{1}{2n^2B}\left[\left(\sum_i \sqrt{b_i}\right)\left(\sum_i \frac{a_i}{\sqrt{b_i}}\right) - \left(\sum_i \sqrt{a_i}\right)^2\right] \geq 0.\end{aligned} \tag{A.47}$$

In this interior case, equality holds if and only if the coefficient vectors $(a_i)_i$ and $(b_i)_i$ are proportional. If the validation gradient norm is $o(1)$ and Eq. A.42 holds for π_α^{IF}, the curvature-optimal design has strictly lower actual risk for all sufficiently large B.

Proof. The general comparison and the sufficient condition for strict actual risk improvement follow by substituting π_α^{IF} into Proposition 4. For interior probabilities, budget calibration gives $\pi_{\alpha,i}^{\mathrm{IF}} = B\sqrt{b_i}/\sum_j \sqrt{b_j}$ and $\pi_{\alpha,i}^\star = B\sqrt{a_i}/\sum_j \sqrt{a_j}$. Consequently,

$$\sum_i \frac{a_i}{\pi_{\alpha,i}^{\mathrm{IF}}} = \frac{1}{B}\left(\sum_i \sqrt{b_i}\right)\left(\sum_i \frac{a_i}{\sqrt{b_i}}\right), \qquad \sum_i \frac{a_i}{\pi_{\alpha,i}^\star} = \frac{1}{B}\left(\sum_i \sqrt{a_i}\right)^2.$$

Subtracting these terms in Eq. 12 proves Eq. A.47. Cauchy–Schwarz gives

$$\left(\sum_i \sqrt{a_i}\right)^2 = \left(\sum_i \sqrt{\frac{a_i}{\sqrt{b_i}}}\sqrt{\sqrt{b_i}}\right)^2 \leq \left(\sum_i \frac{a_i}{\sqrt{b_i}}\right)\left(\sum_i \sqrt{b_i}\right).$$

Because all coefficients are positive, equality requires the two vectors inside the inner product to be proportional, equivalently a_i/b_i constant across i. □

The proportionality criterion concerns the interior case. With active probability bounds, distinct coefficient vectors can yield identical calibrated designs, so equality must be checked at the level of the probabilities. Also, a positive surrogate difference alone does not establish a positive actual-risk difference if the approximation remainder is of the same order. If some $b_i = 0$, the positive-coefficient IF formula does not apply without further conventions. One may instead use a common coefficient floor $\max\{b_i, \tau\}$ with $\tau > 0$ and apply the general comparison to the resulting design; this optimizes the adjusted coefficients. When all $b_i = 0$, this convention gives uniform sampling, whereas the unadjusted IF objective makes every feasible design optimal.

Strict separation with nonzero influence scores The following example separates both baselines from the curvature-optimal design while all scalar IF scores are nonzero. Let n be a multiple of four and use the two-dimensional squared loss $\ell(Z;\theta) = \frac{1}{2}\|\theta - Z\|_2^2$. The training targets Z_i take the four values $(1,0)^\top$, $(-1,0)^\top$, $(1,2)^\top$, and $(-1,-2)^\top$ in equal proportions. The full-data minimizer is

$\hat{\theta} = 0$ and $H_{\mathrm{tr}} = I_2$. Take $B = n/4$, $c = 1/2$, and hence $\alpha = 1/8$. For validation, use the target $-B^{-1/2}e_1$, where $e_1 = (1, 0)^\top$, so that

$$F_{\mathrm{val}}(\theta) = \tfrac{1}{2}\|\theta + B^{-1/2}e_1\|_2^2, \qquad \nabla F_{\mathrm{val}}(0) = B^{-1/2}e_1, \qquad H_{\mathrm{val}} = I_2.$$

The parameter influences are $\mathcal{I}_{\hat{\theta}}(Z_i) = Z_i$. Thus the first two training types have $a_i = 1$, the last two have $a_i = 5$, and every type has $b_i = B^{-1} > 0$. The IF design is uniquely uniform, while the curvature-optimal probabilities are

$$\pi^{\mathrm{IF}}_{\alpha,i} = \pi^{\mathrm{unif}}_i = \frac{1}{4}, \qquad \pi^\star_{\alpha,i} = \begin{cases} \dfrac{1}{2(1+\sqrt{5})}, & a_i = 1, \\ \dfrac{\sqrt{5}}{2(1+\sqrt{5})}, & a_i = 5. \end{cases}$$

Both curvature probabilities lie strictly between $1/8$ and 1. Their score variance is

$$\mathrm{Var}_n(\sqrt{a_i}) = 3 - \left(\frac{1+\sqrt{5}}{2}\right)^2 = \frac{3-\sqrt{5}}{2} > 0.$$

We verify the risk-approximation assumptions rather than assuming them from the quadratic form. For any $\pi \in \mathcal{P}_\alpha(B)$, write $t_\pi = n^{-1}\sum_i W_i/\pi_i$. On a nonempty draw, the exact minimizer is the weighted average $\hat{\theta}_\pi = (\sum_i W_i Z_i/\pi_i)/(\sum_i W_i/\pi_i)$; on the empty draw, return 0. Consequently $\|\hat{\theta}_\pi\|_2 \le \sqrt{5}$, all optimization residuals vanish, and $\overline{H}_\pi = t_\pi I_2$. The full training Hessian has minimum eigenvalue one, and the validation Hessian has operator norm one and Lipschitz constant zero.

For deterministic scalars v_i with $|v_i| \le 2$, set $\xi_i = (W_i/\pi_i - 1)v_i$. Since $\pi_i \ge 1/8$, these independent centered variables have uniformly bounded fourth moments. Expanding their fourth moment and removing terms with an unpaired centered factor gives

$$\mathbb{E}\left(\frac{1}{n}\sum_i \xi_i\right)^4 = \frac{1}{n^4}\left[\sum_i \mathbb{E}\xi_i^4 + 6\sum_{i<j}\mathbb{E}\xi_i^2\mathbb{E}\xi_j^2\right] = \frac{O(n)+O(n^2)}{n^4} = O(n^{-2}) = O(B^{-2}).$$

Choosing $v_i = 1$ bounds $\mathbb{E}|t_\pi - 1|^4$. Choosing each coordinate of Z_i bounds the fourth moment of each coordinate of g^c_π; the inequality $(u_1^2 + u_2^2)^2 \le 2(u_1^4 + u_2^4)$ then gives $\mathbb{E}\|g^c_\pi\|_2^4 = O(B^{-2})$. On $\mathcal{A}_\pi = \{|t_\pi - 1| \le 1/2\}$, the curvature fourth-moment bound follows from the bound for t_π. Markov's inequality and the bounded displacement give

$$\begin{aligned} \mathbb{P}(\mathcal{A}_\pi^c) &\le 16\mathbb{E}|t_\pi - 1|^4 = O(B^{-2}), \\ \mathbb{E}[\|\hat{\theta}_\pi\|_2^2 \mathbf{1}_{\mathcal{A}_\pi^c}] &\le 5\mathbb{P}(\mathcal{A}_\pi^c) = O(B^{-2}), \\ \mathbb{E}[\|\hat{\theta}_\pi\|_2^3 \mathbf{1}_{\mathcal{A}_\pi^c}] &\le 5\sqrt{5}\mathbb{P}(\mathcal{A}_\pi^c) = O(B^{-2}) = O(B^{-3/2}). \end{aligned}$$

These bounds hold throughout the feasible set. The actual-risk minimum is also attained: each nonempty-draw minimizer is continuous in π, its denominator is positive, and the empty-draw minimizer is constant. The finite sum defining expected risk is therefore continuous on the compact feasible set.

All comparison assumptions now hold, with validation gradient norm $B^{-1/2}$. Equation A.45 yields

$$\begin{aligned} \mathcal{R}(\pi^{\mathrm{IF}}_\alpha) - \mathcal{R}(\pi^\star_\alpha) &= \mathcal{R}(\pi^{\mathrm{unif}}) - \mathcal{R}(\pi^\star_\alpha) \\ &= \frac{3-\sqrt{5}}{4B} + O(B^{-3/2}). \end{aligned} \tag{A.48}$$

Theorem 1 with $r_B = B^{-1/2}$ gives $O(B^{-3/2})$ excess risk for the curvature-optimal design. Adding this excess risk to Eq. A.48 gives $(3-\sqrt{5})/(4B) + O(B^{-3/2}) = \Theta(B^{-1})$ excess risk for both baselines relative to the same actual-risk optimizer. Here the scalar IF scores have identical magnitudes, while validation curvature also assigns cost to variation in the second coordinate. The strict difference arises despite all scalar IF scores being nonzero. Appendix A.4.5 supplies a separate nonquadratic cross-entropy example for the uniform comparison.

A.4 SUFFICIENT CONDITIONS AND EXAMPLES

The guarantees in Appendices A.2 and A.3 are conditional on the observed datasets and take Assumptions 1 and 2 as given. This subsection shows when these assumptions hold and how the guarantees transfer to random datasets. It is organized as follows.

- Appendix A.4.1 separates sampling randomness from data randomness and states the resulting guarantee over random datasets (Proposition 6).
- Appendix A.4.2 gives conditions under which the validation gradient vanishes, so that $r_B \to 0$.
- Appendix A.4.3 derives the perturbation and stability bounds from properties of the losses, and verifies all regularity conditions for convex smooth losses with a fixed ridge penalty (Proposition 9).
- Appendices A.4.4 and A.4.5 instantiate this result for frozen-feature softmax and for a binary cross-entropy model in which regularity, positive scores, and a vanishing validation gradient hold jointly (Proposition 11).

A.4.1 CONDITIONAL RISKS AND GUARANTEES OVER RANDOM DATA

Consider sequences of training and validation datasets of sizes n and m, with $B \to \infty$ and $B \le n$. The datasets may now be random. Conditional on the observed datasets and any other quantities fixed before sampling, expectations and probabilities retain the convention in Appendix A.1. We write $\mathbb{E}_W$ and $\mathbb{P}_W$ when only the sampling indicators remain random. The notation $\mathbb{E}_{\mathrm{data}}$ and $\mathbb{P}_{\mathrm{data}}$ refers to the joint law of the datasets; all O_p and o_p statements below use this outer law. Thus $\mathcal{R}(\pi)$ remains a conditional expected empirical validation loss. An outer event specifies which datasets satisfy the required conditions, whereas $\mathcal{A}_\pi$ describes stable retraining under a given dataset and design.

Proposition 6 (Risk comparison over random datasets)**.** *Set $\theta_0 = \hat{\theta}$ and fix $0 < c < 1$ and $\alpha = cB/n$. Suppose that the fitted parameters, risks, and competing designs are well defined and measurable, with finite risks and an attained actual-risk minimum. On data events whose probabilities tend to one, assume the differentiability requirements of Assumption 1, positive definite H, and $a_i > 0$ for all i. Allow the associated bounds to depend on the data. Suppose that $\lambda_{\min}(H)^{-1}$, $\|H_{\mathrm{val}}(\hat{\theta})\|_2$, and L_{val} are $O_p(1)$, and that, separately for $\pi = \pi_\alpha^\star$ and $\pi = \pi_{\mathrm{opt}}$, every quantity in the following display is $O_p(1)$:*

$$\begin{gathered} B^2\mathbb{E}\|g_\pi^{\mathrm{c}}\|_2^4, \qquad B^2\mathbb{E}[\|\overline{H}_\pi - H\|_2^4 \mathbf{1}_{\mathcal{A}_\pi}], \qquad B^2\mathbb{P}(\mathcal{A}_\pi^{\mathrm{c}}), \\ B^2\mathbb{E}[\|\Delta_\pi\|_2^2 \mathbf{1}_{\mathcal{A}_\pi^{\mathrm{c}}}], \qquad B^{3/2}\mathbb{E}[\|\Delta_\pi\|_2^3 \mathbf{1}_{\mathcal{A}_\pi^{\mathrm{c}}}], \\ B^{1/2}\|r_0\|_2, \qquad B(\mathbb{E}\|r_\pi\|_2^4)^{1/4}. \end{gathered} \tag{A.49}$$

The expectations and probabilities inside this display are conditional on the data; they also include subsequent optimizer randomness when present. The residuals are zero under exact fitting and retraining. For any deterministic $r_B > 0$,

$$\begin{aligned} \|\nabla F_{\mathrm{val}}(\hat{\theta})\|_2 = O_p(r_B) \quad \Longrightarrow \quad & 0 \le \mathcal{R}(\pi_\alpha^\star) - \mathcal{R}(\pi_{\mathrm{opt}}) \\ & = O_p(r_B/B + B^{-3/2}). \end{aligned} \tag{A.50}$$

If the validation gradient is $o_p(1)$, the excess risk is $o_p(B^{-1})$.

Proof. Fix an outer failure tolerance $\epsilon > 0$. There are finitely many control quantities in the statement, including both designs. Their $O_p(1)$ bounds and a union bound give deterministic upper bounds on all of them, jointly with probability at least $1 - \epsilon/2$ for all sufficiently large B. Intersect with the data events on which differentiability, positive curvature, and positive scores hold, enlarging the starting index and choosing the preceding tolerances smaller if needed. On this intersection, the proof of Proposition 3 uses only these bounds. In particular, with d_0 from Eq. A.31 and $\beta_\pi = e_\pi - d_0$ as in that proof, its moment inequalities give

$$\|d_0\|_2 \le CB^{-1/2}, \quad \mathbb{E}\|\beta_\pi\|_2^2 \le CB^{-2}, \quad \mathbb{E}\|\Delta_\pi^{\mathrm{lin}}\|_2^2 \le CB^{-1}, \quad \mathbb{E}\|\Delta_\pi\|_2^3 \le CB^{-3/2},$$

where C is deterministic and depends only on the chosen bounds. Indeed, the stable-event displacement inequality uses inverse curvature and the three gradient terms; its fourth moment is bounded by a constant times $B^{-2} + B^{-2} + B^{-4}$. The squared error bound then uses a terminal-residual term of order B^{-2} and a product of two square-root fourth moments, each of order B^{-1}. The exceptional error is controlled by the second displacement moment and the product of a gradient square-root fourth moment and a square-root failure probability, again $B^{-1}B^{-1} = B^{-2}$. These are precisely the inequalities in Eqs. A.35–A.37.

Substitution into Eq. A.38 bounds each corrected risk difference by a constant times

$$\|\nabla F_{\mathrm{val}}(\hat{\theta})\|_2 B^{-1} + B^{-1/2}B^{-1} + B^{-2} + B^{-3/2} \le \|\nabla F_{\mathrm{val}}(\hat{\theta})\|_2 B^{-1} + 3B^{-3/2}$$

for $B \ge 1$. The common shift C_B cancels between designs and the surrogate difference is nonpositive, exactly as in the proof of Eq. A.40. Its constant can therefore be chosen uniformly on this data event. No bound on the validation gradient was used in selecting that constant.

The assumption $\|\nabla F_{\mathrm{val}}(\hat{\theta})\|_2 = O_p(r_B)$ gives a deterministic bound on its ratio to r_B outside a data event of probability at most $\epsilon/2$ for all sufficiently large B. Intersecting the two events proves Eq. A.50: the two contributions are $O_p(r_B)B^{-1} = O_p(r_B/B)$ and $O_p(1)B^{-3/2}$. For a gradient of order $o_p(1)$, multiply the conditional bound by B. On each event with the chosen deterministic controls, its right side is $C[\|\nabla F_{\mathrm{val}}(\hat{\theta})\|_2 + B^{-1/2}]$, which converges to zero in probability. The excluded probability can be made arbitrarily small by choosing ϵ, proving the claimed $o_p(B^{-1})$ rate. □

The normalized bounds above state what must be transferred to the outer law; the following subsections give sufficient conditions for them. A stronger premise, namely that one set of deterministic constants works on data events whose probabilities tend to one, implies these $O_p(1)$ controls. The converse need not hold, because tightness allows the constants to depend on the prescribed failure tolerance. In either formulation, $\mathcal{R}(\pi)$ remains a conditional risk rather than an expectation over the datasets.

A.4.2 Sufficient conditions for a vanishing validation gradient

In this subsection the training and validation samples are independent, and each sample is independently and identically distributed under its own population law. The loss and any feature map are fixed before these datasets are drawn. Write $\mathbb{E}_{\mathrm{tr}}$ and $\mathbb{E}_{\mathrm{val}}$ for expectation over one population training or validation example. Population objectives are assumed differentiable with gradients given by these expectations. They serve only to verify the conditions on the empirical validation gradient. Let $\theta^\circ_{\mathrm{tr}}$ be a deterministic stationary point of the population counterpart of the actual training objective, including any deterministic training penalty. It may vary along the problem sequence. Define the population mismatch

$$\delta_{\mathrm{tv}} := \|\mathbb{E}_{\mathrm{val}} \nabla \ell(Z; \theta^\circ_{\mathrm{tr}})\|_2 . \tag{A.51}$$

Proposition 7 (Statistical control of the validation gradient)**.** *Suppose $\|\hat{\theta} - \theta^\circ_{\mathrm{tr}}\|_2 = O_p(r_n)$ for a deterministic $r_n > 0$, and the total gradient variance satisfies*

$$\mathbb{E}_{\mathrm{val}} \|\nabla \ell(Z; \theta^\circ_{\mathrm{tr}}) - \mathbb{E}_{\mathrm{val}} \nabla \ell(Z; \theta^\circ_{\mathrm{tr}})\|_2^2 = O(1). \tag{A.52}$$

With probability tending to one, assume F_{val} is twice continuously differentiable on a neighborhood of the segment joining $\theta^\circ_{\mathrm{tr}}$ and $\hat{\theta}$, and

$$\sup_{0 \le t \le 1} \left\| H_{\mathrm{val}} \left[\theta^\circ_{\mathrm{tr}} + t(\hat{\theta} - \theta^\circ_{\mathrm{tr}}) \right] \right\|_2 = O_p(1).$$

Then

$$\|\nabla F_{\mathrm{val}}(\hat{\theta})\|_2 = O_p(\delta_{\mathrm{tv}} + r_n + m^{-1/2}). \tag{A.53}$$

In particular, $\delta_{\mathrm{tv}} \to 0$, $r_n \to 0$, and $m \to \infty$ imply $\|\nabla F_{\mathrm{val}}(\hat{\theta})\|_2 = o_p(1)$.

Proof. Add and subtract the empirical and population validation gradients at the deterministic reference point:

$$\begin{aligned}\nabla F_{\mathrm{val}}(\hat{\theta}) = \mathbb{E}_{\mathrm{val}}\nabla\ell(Z;\theta^\circ_{\mathrm{tr}}) \\ + [\nabla F_{\mathrm{val}}(\theta^\circ_{\mathrm{tr}}) - \mathbb{E}_{\mathrm{val}}\nabla\ell(Z;\theta^\circ_{\mathrm{tr}})] \\ + \left[\nabla F_{\mathrm{val}}(\hat{\theta}) - \nabla F_{\mathrm{val}}(\theta^\circ_{\mathrm{tr}})\right].\end{aligned}$$

The first term has norm δ_{tv}. The second is an average of independent centered vectors. Expanding its squared norm, every cross term has expectation zero, so

$$\begin{aligned}&\mathbb{E}_{\mathrm{data}}\left\|\nabla F_{\mathrm{val}}(\theta^\circ_{\mathrm{tr}}) - \mathbb{E}_{\mathrm{val}}\nabla\ell(Z;\theta^\circ_{\mathrm{tr}})\right\|_2^2 \\ &\quad= \frac{1}{m}\mathbb{E}_{\mathrm{val}}\left\|\nabla\ell(Z;\theta^\circ_{\mathrm{tr}}) - \mathbb{E}_{\mathrm{val}}\nabla\ell(Z;\theta^\circ_{\mathrm{tr}})\right\|_2^2 = O(m^{-1}).\end{aligned}$$

Markov's inequality applied to this squared norm gives $O_p(m^{-1/2})$ for the norm. For the last term, the fundamental theorem of calculus on the stated segment gives

$$\begin{aligned}&\nabla F_{\mathrm{val}}(\hat{\theta}) - \nabla F_{\mathrm{val}}(\theta^\circ_{\mathrm{tr}}) \\ &\quad= \int_0^1 H_{\mathrm{val}}\left[\theta^\circ_{\mathrm{tr}} + t(\hat{\theta} - \theta^\circ_{\mathrm{tr}})\right](\hat{\theta} - \theta^\circ_{\mathrm{tr}})\,\mathrm{d}t.\end{aligned}$$

Its norm is at most the displayed supremum times the estimation error, of order $O_p(1)O_p(r_n) = O_p(r_n)$. The exceptional data probability tends to zero, so it does not affect this probability order. Summing the three terms proves Eq. A.53. When its deterministic scale tends to zero, tightness of the ratio implies convergence to zero in probability. □

Obtaining the training estimation rate A sufficient route to $r_n = n^{-1/2}$ uses positive curvature and population stationarity. First consider the unpenalized objective. Suppose the empirical training Hessian is at least μI on the segment from $\theta^\circ_{\mathrm{tr}}$ to $\hat{\theta}$, on data events of probability tending to one, with $\mu^{-1} = O_p(1)$. Integrating its Hessian and applying Cauchy–Schwarz yields

$$\begin{aligned}\mu\|\hat{\theta} - \theta^\circ_{\mathrm{tr}}\|_2^2 &\le (\hat{\theta} - \theta^\circ_{\mathrm{tr}})^\top[\nabla F_{\mathrm{tr}}(\hat{\theta}) - \nabla F_{\mathrm{tr}}(\theta^\circ_{\mathrm{tr}})] \\ &\le \|\hat{\theta} - \theta^\circ_{\mathrm{tr}}\|_2[\|\nabla F_{\mathrm{tr}}(\hat{\theta})\|_2 + \|\nabla F_{\mathrm{tr}}(\theta^\circ_{\mathrm{tr}})\|_2].\end{aligned}$$

Divide by the displacement norm when it is nonzero; the resulting bound also holds when it is zero. Population stationarity and a uniformly bounded total training score variance give

$$\mathbb{E}_{\mathrm{data}}\|\nabla F_{\mathrm{tr}}(\theta^\circ_{\mathrm{tr}})\|_2^2 = \frac{1}{n}\mathbb{E}_{\mathrm{tr}}\left\|\nabla\ell(Z;\theta^\circ_{\mathrm{tr}}) - \mathbb{E}_{\mathrm{tr}}\nabla\ell(Z;\theta^\circ_{\mathrm{tr}})\right\|_2^2 = O(n^{-1}).$$

Thus exact fitting, or a training gradient residual of order $O_p(n^{-1/2})$, yields an estimation error of order $O_p(1)[O_p(n^{-1/2}) + O_p(n^{-1/2})] = O_p(n^{-1/2})$. For a deterministic penalty, apply the same integrated-Hessian inequality to the penalized objective and define $\theta^\circ_{\mathrm{tr}}$ using that objective. Its gradient at the population point is still the centered empirical average of the data scores: the population mean and the penalty gradient cancel. For ridge, the curvature lower bound is at least λ whenever the data losses are convex. This gives the required estimation rate without an additional consistency assumption in that setting.

Alignment and sample sizes The mismatch condition concerns the actual training target. For fixed population objectives and a fixed population training point, the mismatch is constant, so $\delta_{\mathrm{tv}} \to 0$ requires $\delta_{\mathrm{tv}} = 0$. Even identical training and validation laws need not give this when training uses fixed ridge and validation is unpenalized: population training stationarity then gives $\mathbb{E}_{\mathrm{val}}\nabla\ell(Z;\theta^\circ_{\mathrm{tr}}) = -\lambda\theta^\circ_{\mathrm{tr}}$. With zero mismatch and the root-n rate, the gradient has order $O_p(n^{-1/2} + m^{-1/2})$. When the regularity conditions in Proposition 6 also hold for the same training objective and reference, the excess risk is

$$O_p\left(\frac{1}{B\sqrt{n}} + \frac{1}{B\sqrt{m}} + B^{-3/2}\right). \tag{A.54}$$

Both sample sizes tending to infinity gives $o_p(B^{-1})$. If $m \gtrsim B$, then $n \geq B$ bounds each of the first two terms by a constant times $B^{-3/2}$, giving $O_p(B^{-3/2})$. For nonzero mismatch, add $O_p(\delta_{\mathrm{tv}}/B)$; $\delta_{\mathrm{tv}} = O(B^{-1/2})$ suffices for the faster rate. The variance assumption is on the whole gradient vector. If its total variance grows with dimension, the noise rate is its square root divided by $\sqrt{m}$, with the corresponding training factor in the estimation rate; coordinatewise bounds alone do not give dimension-independent rates.

A.4.3 Sufficient conditions for perturbation control and stability

We now derive the conditional controls used above from loss derivatives, curvature, and properties of the training outputs. The perturbation bounds apply to a general fixed reference. The subsequent local arguments specialize to an exact full-data fit, and a final result verifies all regularity conditions for convex smooth losses with fixed ridge. These sufficient conditions do not require a particular softmax or logistic form.

Perturbation moments under a sampling budget Suppose there are fixed constants $c \in (0,1)$ and $G, K < \infty$, independent of B, such that $\pi_i \geq cB/n$, $\|\nabla_\theta \ell(Z_i;\theta_0)\|_2 \leq G$, and $\|\nabla^2_\theta \ell(Z_i;\theta_0)\|_F \leq K$ for every i. Allow $\sum_i \pi_i \leq B$; the moment bounds below use only the probability floor and do not require budget equality. Within the perturbation calculations below, write $s_i := \nabla_\theta \ell(Z_i;\theta_0)$, $H_i := \nabla^2_\theta \ell(Z_i;\theta_0)$, and $\delta_i := W_i/\pi_i - 1$. Define $E_\pi := \nabla^2_\theta F_{\mathrm{tr},\pi}(\theta_0; W) - H_{\mathrm{tr}}(\theta_0)$. Centering, without a stationarity assumption, gives

$$g^{\mathrm{c}}_\pi = \frac{1}{n}\sum_i \delta_i s_i, \qquad E_\pi = \frac{1}{n}\sum_i \delta_i H_i. \tag{A.55}$$

The moments in Eqs. A.16 and A.17 imply

$$\mathbb{E}\|g^{\mathrm{c}}_\pi\|_2^2 = \frac{1}{n^2}\sum_{i,j}\mathbb{E}[\delta_i\delta_j] s_i^\top s_j = \frac{1}{n^2}\sum_i \left(\frac{1}{\pi_i} - 1\right)\|s_i\|_2^2, \tag{A.56}$$

$$\begin{aligned}\mathbb{E}\|E_\pi\|_2^2 \leq \mathbb{E}\|E_\pi\|_F^2 &= \frac{1}{n^2}\sum_{i,j}\mathbb{E}[\delta_i\delta_j]\,\mathrm{tr}(H_i^\top H_j)\\ &= \frac{1}{n^2}\sum_i \left(\frac{1}{\pi_i} - 1\right)\|H_i\|_F^2.\end{aligned} \tag{A.57}$$

These expressions can be evaluated using the actual probabilities. For the stated budget bounds, $\|s_i\|_2 \leq G$ and $\pi_i^{-1} \leq n/(cB)$ give $\mathbb{E}\|g^{\mathrm{c}}_\pi\|_2^2 \leq n^{-2} n[n/(cB)]G^2 = G^2/(cB)$. The identical scalar bound on the sum in Eq. A.57, with $\|H_i\|_F \leq K$, gives $\mathbb{E}\|E_\pi\|_F^2 \leq K^2/(cB)$.

Fourth moments We first derive the needed inequality for independent centered vectors ξ_i in a finite-dimensional real inner-product space. Within this argument let $v := \sum_i \mathbb{E}\|\xi_i\|^2$ and suppose $\|\xi_i\| \leq b$ almost surely. We expand $\|\sum_i \xi_i\|^2 = \sum_i \|\xi_i\|^2 + 2\sum_{i<j}\langle \xi_i, \xi_j\rangle$ and square. Every term with an index occurring only once has expectation zero by independence and centering. The remaining terms are

$$\begin{aligned}\mathbb{E}\left\|\sum_i \xi_i\right\|^4 &= \sum_i \mathbb{E}\|\xi_i\|^4 + 2\sum_{i<j}\mathbb{E}\|\xi_i\|^2\,\mathbb{E}\|\xi_j\|^2 + 4\sum_{i<j}\mathbb{E}\langle\xi_i,\xi_j\rangle^2\\ &\leq b^2 v + 6\sum_{i<j}\mathbb{E}\|\xi_i\|^2\,\mathbb{E}\|\xi_j\|^2 \leq b^2 v + 3v^2.\end{aligned} \tag{A.58}$$

The first inequality uses $\|\xi_i\|^4 \leq b^2\|\xi_i\|^2$ and $\langle\xi_i,\xi_j\rangle^2 \leq \|\xi_i\|^2\|\xi_j\|^2$ together with independence. The last inequality follows from $v^2 = \sum_i(\mathbb{E}\|\xi_i\|^2)^2 + 2\sum_{i<j}\mathbb{E}\|\xi_i\|^2\mathbb{E}\|\xi_j\|^2$.

For the gradient sum, take $\xi_i = \delta_i s_i/n$. Since $|\delta_i| \leq 1/\pi_i$, $\|\xi_i\|_2 \leq G/(n\pi_i) \leq G/(cB)$. The variance sum is at most $G^2/(cB)$. For the Hessian sum, use $\xi_i = \delta_i H_i/n$ in the space of matrices with the Frobenius inner product. Its summands have norm at most $K/(cB)$ and variance sum at

most $K^2/(cB)$. Substitution into Eq. A.58 gives

$$\mathbb{E}\|g_\pi^{\mathrm{c}}\|_2^4 \le \frac{G^4}{c^3B^3} + \frac{3G^4}{c^2B^2} = O(B^{-2}), \tag{A.59}$$

$$\mathbb{E}\|E_\pi\|_2^4 \le \mathbb{E}\|E_\pi\|_F^4 \le \frac{K^4}{c^3B^3} + \frac{3K^4}{c^2B^2} = O(B^{-2}). \tag{A.60}$$

The rates assume fixed c, G, K and $B \to \infty$. If the parameter dimension grows, the dependence of K on that dimension must be retained; the Frobenius bound is not asserted to be dimension independent. For any fixed $u > 0$, Markov's inequality also gives

$$\mathbb{P}\{\|g_\pi^{\mathrm{c}}\|_2 > u \text{ or } \|E_\pi\|_2 > u\} \le u^{-4}\left(\mathbb{E}\|g_\pi^{\mathrm{c}}\|_2^4 + \mathbb{E}\|E_\pi\|_2^4\right) = O(B^{-2}).$$

Sampled smoothness Suppose the Hessian of each example loss is L_ℓ-Lipschitz on the ball of radius ρ around θ_0. The nonnegative HT weights imply, for any θ, θ' in this ball,

$$\|\nabla_\theta^2 F_{\mathrm{tr},\pi}(\theta; W) - \nabla_\theta^2 F_{\mathrm{tr},\pi}(\theta'; W)\|_2 \le L_\ell \left(\frac{1}{n}\sum_i \frac{W_i}{\pi_i}\right)\|\theta - \theta'\|_2.$$

Apply Eq. A.58 to the scalar summands $\xi_i = \delta_i/n$. Their absolute values are at most $1/(cB)$ and their variance sum is at most $1/(cB)$, so

$$\mathbb{E}\left|\frac{1}{n}\sum_i \frac{W_i}{\pi_i} - 1\right|^4 \le \frac{1}{c^3B^3} + \frac{3}{c^2B^2} = O(B^{-2}). \tag{A.61}$$

In particular, the total HT weight is at most two except on an event of probability $O(B^{-2})$, and on that event's complement the sampled Hessian has Lipschitz constant at most $2L_\ell$. These bounds control sampling perturbations and smoothness; localization of the training output and its exceptional moments remain additional requirements.

Empirical derivative moments and random reference points The uniform bounds G, K above give one sufficient condition. The fourth-moment expansion also applies without bounding each summand by a common constant: before that bound is used, it gives

$$\mathbb{E}\left\|\sum_i \xi_i\right\|^4 \le \sum_i \mathbb{E}\|\xi_i\|^4 + 3\left(\sum_i \mathbb{E}\|\xi_i\|^2\right)^2.$$

Because $|\delta_i| \le \pi_i^{-1}$ and $\mathbb{E}_W\delta_i^2 = \pi_i^{-1} - 1 \le \pi_i^{-1}$, $\mathbb{E}_W\delta_i^4 \le \pi_i^{-2}\mathbb{E}_W\delta_i^2 \le \pi_i^{-3}$. Substituting $\xi_i = \delta_i s_i/n$ and $\pi_i \ge cB/n$ gives

$$\mathbb{E}_W\|g_\pi^{\mathrm{c}}\|_2^4 \le \frac{1}{c^3B^3}\frac{1}{n}\sum_i \|s_i\|_2^4 + \frac{3}{c^2B^2}\left(\frac{1}{n}\sum_i \|s_i\|_2^2\right)^2, \tag{A.62}$$

$$\mathbb{E}_W\|E_\pi\|_2^4 \le \frac{1}{c^3B^3}\frac{1}{n}\sum_i \|H_i\|_F^4 + \frac{3}{c^2B^2}\left(\frac{1}{n}\sum_i \|H_i\|_F^2\right)^2. \tag{A.63}$$

For the second inequality, use the Frobenius inner product and then $\|E_\pi\|_2 \le \|E_\pi\|_F$. If the empirical fourth moments are $O_p(1)$, Cauchy–Schwarz bounds each squared empirical second moment by the corresponding fourth moment. Multiplying either inequality by B^2 gives $B^{-1}O_p(1) + O_p(1) = O_p(1)$.

A sufficient population condition is a fourth-power derivative envelope with uniformly bounded population expectation along the sequence, on a deterministic neighborhood that contains θ_0 with probability tending to one. The empirical mean of that envelope is $O_p(1)$ by Markov's inequality, and bounds the empirical derivative moments even when $\theta_0 = \hat{\theta}$ depends on all training observations. The fitted scores themselves need not be independent. For sampled smoothness, the preceding uniform local Lipschitz bound and HT-mass calculation remain in force; moment control at the reference alone does not control derivatives along an unlocalized retraining segment.

Positive reference curvature can also be checked at a deterministic population point. Suppose the population training Hessian there is at least μI for fixed $\mu > 0$, its empirical average differs from its expectation by $o_p(1)$ in operator norm, and the empirical Hessian has Lipschitz constant $O_p(1)$ between that point and a consistent fit. The triangle inequality gives

$$\|H_{\mathrm{tr}}(\hat{\theta}) - \mathbb{E}_{\mathrm{tr}}\nabla^2\ell(Z;\theta^\circ_{\mathrm{tr}})\|_2 \le \|H_{\mathrm{tr}}(\theta^\circ_{\mathrm{tr}}) - \mathbb{E}_{\mathrm{tr}}\nabla^2\ell(Z;\theta^\circ_{\mathrm{tr}})\|_2 + O_p(1)\|\hat{\theta} - \theta^\circ_{\mathrm{tr}}\|_2 = o_p(1) + O_p(1)o_p(1) = o_p(1).$$

The empirical minimum eigenvalue is then at least $\mu/2$ with probability tending to one. For example, bounded total Frobenius variance at the deterministic point gives an empirical Hessian error of order $O_p(n^{-1/2})$ by the centered-average second-moment calculation. For a smooth deterministic penalty, include its Hessian in both training Hessians and in the population curvature condition. Fixed ridge instead supplies a global lower bound directly.

Risk approximation at a general perturbation scale The local analysis below, up to the frozen-feature softmax paragraph, uses the exact-stationarity setting of Appendix A.2.4. Consider a fixed design π and the displacement Δ_π, error e_π, and risk bound $\varepsilon(\pi)$ defined above. Here $g_\pi = g^{\mathrm{c}}_\pi$ because $r_0 = 0$. The sampled gradient and Hessian perturbation at the full-data solution are

$$\begin{aligned} g_\pi &:= \nabla_\theta F_{\mathrm{tr},\pi}(\hat{\theta}; W), \\ E_\pi &:= \nabla^2_\theta F_{\mathrm{tr},\pi}(\hat{\theta}; W) - H_{\mathrm{tr}}(\hat{\theta}). \end{aligned} \tag{A.64}$$

Thus $\Delta^{\mathrm{lin}}_\pi = -H_{\mathrm{tr}}(\hat{\theta})^{-1} g_\pi$. We use an auxiliary localization event $\mathcal{G}_\pi$ to distinguish local retraining from exceptional draws. The convex localization argument below constructs an explicit choice and relates it to the event in Eq. A.5 for this design.

Proposition 8 (Risk accuracy at a general perturbation scale)**.** *Fix a sequence of problems and designs π, each chosen before sampling, with perturbation scale $q \downarrow 0$. Suppose the assumptions of Proposition 2 hold, $H_{\mathrm{tr}}(\hat{\theta}) \succeq \mu I$ for a fixed $\mu > 0$, and $\|H_{\mathrm{val}}(\hat{\theta})\|_2$ and L_{val} remain bounded. Assume the training losses are twice continuously differentiable near the retraining segments on $\mathcal{G}_\pi$. For fixed $\rho > 0$ and $L_{\mathrm{tr}} \ge 0$, suppose that on this event $\|\Delta_\pi\|_2 \le \rho$, the sampled training Hessian is L_{tr}-Lipschitz along the segment, and*

$$\|E_\pi\|_2 + \frac{L_{\mathrm{tr}}\rho}{2} \le \frac{\mu}{2}. \tag{A.65}$$

Assume also

$$\mathbb{E}_W\|g_\pi\|_2^4 = O(q^4), \qquad \mathbb{E}_W\|E_\pi\|_2^4 = O(q^4), \tag{A.66}$$

and

$$\begin{aligned} \mathbb{P}_W(\mathcal{G}^c_\pi) &= O(q^4), \\ \mathbb{E}_W\left[\|\Delta_\pi\|_2^2 \mathbf{1}_{\mathcal{G}^c_\pi}\right] &= O(q^4), \\ \mathbb{E}_W\left[\|\Delta_\pi\|_2^3 \mathbf{1}_{\mathcal{G}^c_\pi}\right] &= O(q^3). \end{aligned} \tag{A.67}$$

All rate constants are independent of q. Then

$$\mathbb{E}_W\|e_\pi\|_2^2 = O(q^4), \qquad \varepsilon(\pi) = O\Big(\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2 q^2 + q^3\Big). \tag{A.68}$$

Proof. All expectations below are under the fixed design, conditional as above. Within this proof, write $H := H_{\mathrm{tr}}(\hat{\theta})$. Full-data stationarity and the influence definition give

$$g_\pi = \frac{1}{n}\sum_i \left(\frac{W_i}{\pi_i} - 1\right)\nabla_\theta \ell(Z_i;\hat{\theta}), \qquad H\Delta^{\mathrm{lin}}_\pi = -g_\pi. \tag{A.69}$$

Local retraining error Fix a draw in $\mathcal{G}_\pi$. Differentiability at the selected unconstrained minimizer gives $\nabla_\theta F_{\mathrm{tr},\pi}(\hat{\theta} + \Delta_\pi; W) = 0$. The fundamental theorem of calculus along the retraining segment gives

$$0 = g_\pi + \overline{H}_\pi \Delta_\pi, \qquad \overline{H}_\pi := \int_0^1 \nabla^2_\theta F_{\mathrm{tr},\pi}(\hat{\theta} + t\Delta_\pi; W)\,\mathrm{d}t. \tag{A.70}$$

At $t = 0$, the sampled Hessian equals $H + E_\pi$. Its Lipschitz property therefore gives

$$\begin{aligned}\|\overline{H}_\pi - H\|_2 &\le \|E_\pi\|_2 + \int_0^1 L_{\mathrm{tr}} t \|\Delta_\pi\|_2 \,\mathrm{d}t \\ &= \|E_\pi\|_2 + \frac{L_{\mathrm{tr}}}{2}\|\Delta_\pi\|_2 \le \frac{\mu}{2}. \end{aligned} \tag{A.71}$$

For every vector u, $u^\top \overline{H}_\pi u \ge \mu\|u\|_2^2 - \|\overline{H}_\pi - H\|_2\|u\|_2^2 \ge (\mu/2)\|u\|_2^2$. Thus $\overline{H}_\pi$ is invertible with $\|\overline{H}_\pi^{-1}\|_2 \le 2/\mu$, and Eq. A.70 yields

$$\|\Delta_\pi\|_2 \le \frac{2}{\mu}\|g_\pi\|_2 \quad \text{on } \mathcal{G}_\pi. \tag{A.72}$$

Subtracting $H\Delta_\pi^{\mathrm{lin}} = -g_\pi$ from the averaged stationarity equation gives $He_\pi = (H - \overline{H}_\pi)\Delta_\pi$. Using the first bound in Eq. A.71, $\|H^{-1}\|_2 \le 1/\mu$, and Eq. A.72, we obtain

$$\begin{aligned}\|e_\pi\|_2 &\le \frac{1}{\mu}\left(\|E_\pi\|_2 + \frac{L_{\mathrm{tr}}}{2}\|\Delta_\pi\|_2\right)\|\Delta_\pi\|_2 \\ &\le \frac{2}{\mu^2}\|E_\pi\|_2\|g_\pi\|_2 + \frac{2L_{\mathrm{tr}}}{\mu^3}\|g_\pi\|_2^2 \quad \text{on } \mathcal{G}_\pi. \end{aligned} \tag{A.73}$$

The two contributions are the product of curvature and gradient perturbations, and the nonlinear training remainder.

Expected squared linearization error Squaring Eq. A.73, using $(a+b)^2 \le 2a^2 + 2b^2$, and then applying Cauchy–Schwarz to the product of squared norms gives

$$\begin{aligned}&\mathbb{E}_W\left[\|e_\pi\|_2^2 \mathbf{1}_{\mathcal{G}_\pi}\right] \\ &\le \frac{8}{\mu^4}\mathbb{E}_W[\|E_\pi\|_2^2\|g_\pi\|_2^2] + \frac{8L_{\mathrm{tr}}^2}{\mu^6}\mathbb{E}_W\|g_\pi\|_2^4 \\ &\le \frac{8}{\mu^4}\left(\mathbb{E}_W\|E_\pi\|_2^4\right)^{1/2}\left(\mathbb{E}_W\|g_\pi\|_2^4\right)^{1/2} + \frac{8L_{\mathrm{tr}}^2}{\mu^6}\mathbb{E}_W\|g_\pi\|_2^4 \\ &= O(q^2)O(q^2) + O(q^4) = O(q^4). \end{aligned} \tag{A.74}$$

The factors have orders $O(q^2)$ because each is the square root of an $O(q^4)$ fourth moment. This calculation does not require independence between g_π and E_π.

On $\mathcal{G}_\pi^c$, the identity $e_\pi = \Delta_\pi + H^{-1}g_\pi$ remains valid. The squared triangle inequality and Cauchy–Schwarz give

$$\begin{aligned}&\mathbb{E}_W\left[\|e_\pi\|_2^2 \mathbf{1}_{\mathcal{G}_\pi^c}\right] \\ &\le 2\mathbb{E}_W\left[\|\Delta_\pi\|_2^2 \mathbf{1}_{\mathcal{G}_\pi^c}\right] + \frac{2}{\mu^2}\mathbb{E}_W\left[\|g_\pi\|_2^2 \mathbf{1}_{\mathcal{G}_\pi^c}\right] \\ &\le 2\mathbb{E}_W\left[\|\Delta_\pi\|_2^2 \mathbf{1}_{\mathcal{G}_\pi^c}\right] + \frac{2}{\mu^2}\left(\mathbb{E}_W\|g_\pi\|_2^4\right)^{1/2}\mathbb{P}_W(\mathcal{G}_\pi^c)^{1/2} \\ &= O(q^4) + O(q^2)O(q^2) = O(q^4). \end{aligned} \tag{A.75}$$

Here the first term uses the assumed exceptional second moment; the product uses the fourth-moment bound and the $O(q^4)$ failure probability. Adding Eqs. A.74 and A.75 gives

$$\mathbb{E}_W\|e_\pi\|_2^2 = O(q^4), \qquad \|\mathbb{E}_W e_\pi\|_2 \le \left(\mathbb{E}_W\|e_\pi\|_2^2\right)^{1/2} = O(q^2). \tag{A.76}$$

Remaining moments and risk transfer The identity $\Delta_\pi^{\mathrm{lin}} = -H^{-1}g_\pi$ gives

$$\mathbb{E}_W\|\Delta_\pi^{\mathrm{lin}}\|_2^2 \le \mu^{-2}\mathbb{E}_W\|g_\pi\|_2^2 \le \mu^{-2}\left(\mathbb{E}_W\|g_\pi\|_2^4\right)^{1/2} = O(q^2). \tag{A.77}$$

On $\mathcal{G}_\pi$, Eq. A.72 controls the actual displacement. On its complement we use the assumed third moment. Hölder's inequality gives $\mathbb{E}_W\|g_\pi\|_2^3 \le (\mathbb{E}_W\|g_\pi\|_2^4)^{3/4}$, so

$$\begin{aligned}\mathbb{E}_W\|\Delta_\pi\|_2^3 &\le \left(\frac{2}{\mu}\right)^3\left(\mathbb{E}_W\|g_\pi\|_2^4\right)^{3/4} + \mathbb{E}_W\left[\|\Delta_\pi\|_2^3 \mathbf{1}_{\mathcal{G}_\pi^c}\right] \\ &= O(q^3) + O(q^3) = O(q^3). \end{aligned} \tag{A.78}$$

In the four terms defining $\varepsilon(\pi)$, the bias factor is $O(q^2)$; the two square-root moments in the cross term are $O(q)$ and $O(q^2)$; the squared error is $O(q^4)$; and the actual third moment is $O(q^3)$. Because the validation Hessian norm and L_{val} are bounded,

$$\begin{aligned}\varepsilon(\pi) &= \|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2\, O(q^2) + O(1)O(q)O(q^2) + O(1)O(q^4) + O(1)O(q^3) \\ &= O\Big(\|\nabla_\theta F_{\mathrm{val}}(\hat{\theta})\|_2 q^2 + q^3\Big). \end{aligned} \tag{A.79}$$

The last step uses $q^4 \le q^3$ for $0 < q \le 1$. This proves Eq. A.68. □

A sufficient bound on exceptional draws If $\|\Delta_\pi\|_2 \le D$ for a fixed finite D on every draw, then

$$\mathbb{E}_W\big[\|\Delta_\pi\|_2^k \mathbf{1}_{\mathcal{G}_\pi^c}\big] \le D^k \mathbb{P}_W(\mathcal{G}_\pi^c) = O(q^4), \qquad k = 2, 3.$$

For $k = 3$ this implies the required $O(q^3)$ bound because $q^4 \le q^3$ for small q. The tail conditions in the proposition allow more general unbounded displacements. A probability tending to one without a rate, or a local expansion valid only on that event, does not by itself supply the moment bounds used in the proof.

Localization for convex training losses For convex training losses, a local stationary point is a global minimizer, which allows localization to follow from perturbation control. Suppose $H_{\mathrm{tr}}(\hat{\theta}) \succeq \mu I$ and the moment and smoothness conditions of the perturbation calculation above hold. Fix $\rho > 0$ small enough that $2L_\ell \rho \le \mu/4$. An explicit choice of the auxiliary localization event is

$$\mathcal{G}_\pi := \left\{ \|g_\pi\|_2 \le \frac{\mu\rho}{4},\ \|E_\pi\|_2 \le \frac{\mu}{4},\ \frac{1}{n}\sum_i \frac{W_i}{\pi_i} \le 2 \right\}. \tag{A.80}$$

For the first two conditions, the fourth-moment bounds and Markov's inequality give failure probabilities $O(B^{-2})$ at the fixed thresholds $\mu\rho/4$ and $\mu/4$. For the third, a violation implies $|n^{-1}\sum_i W_i/\pi_i - 1| > 1$; Eq. A.61 gives the same rate. The union bound yields $\mathbb{P}_W(\mathcal{G}_\pi^c) \le O(B^{-2}) + O(B^{-2}) + O(B^{-2}) = O(B^{-2})$.

On $\mathcal{G}_\pi$, the sampled Hessian is L_{tr}-Lipschitz on the ball with $L_{\mathrm{tr}} = 2L_\ell$. For every θ in that ball,

$$\|\nabla_\theta^2 F_{\mathrm{tr},\pi}(\theta; W) - H_{\mathrm{tr}}(\hat{\theta})\|_2 \le \|E_\pi\|_2 + L_{\mathrm{tr}}\|\theta - \hat{\theta}\|_2 \le \frac{\mu}{4} + \frac{\mu}{4} = \frac{\mu}{2}.$$

For every vector u, this bound gives $u^\top \nabla_\theta^2 F_{\mathrm{tr},\pi}(\theta; W) u \ge (\mu/2)\|u\|_2^2$. For any two points $\theta, \theta + h$ in the ball, the integral Taylor formula then gives

$$F_{\mathrm{tr},\pi}(\theta + h; W) - F_{\mathrm{tr},\pi}(\theta; W) - \nabla_\theta F_{\mathrm{tr},\pi}(\theta; W)^\top h \ge \int_0^1 (1 - t)\frac{\mu}{2}\|h\|_2^2 \,\mathrm{d}t = \frac{\mu}{4}\|h\|_2^2.$$

This is $\mu/2$-strong convexity on the ball. For a boundary displacement h with $\|h\|_2 = \rho$, integrating the gradient gives

$$\begin{aligned} h^\top \nabla_\theta F_{\mathrm{tr},\pi}(\hat{\theta} + h; W) &= h^\top g_\pi + \int_0^1 h^\top \nabla_\theta^2 F_{\mathrm{tr},\pi}(\hat{\theta} + th; W) h \,\mathrm{d}t \\ &\ge -\rho\|g_\pi\|_2 + \frac{\mu}{2}\rho^2 \ge -\frac{\mu\rho^2}{4} + \frac{\mu\rho^2}{2} = \frac{\mu\rho^2}{4} > 0. \end{aligned}$$

The continuous objective attains a minimum on the closed ball. At a boundary point, the directional derivative toward the center is the negative of the displayed positive quantity; a sufficiently small step into the ball decreases the objective. No boundary point can therefore minimize it on the ball. The minimizer lies in the interior and has zero gradient. Global convexity implies that this point is a global minimizer: the convex gradient inequality bounds the objective at any other point from below by its value at the stationary point. If there were another global minimizer, convexity would make the segment between them consist of global minimizers. A nontrivial part of that segment lies inside the ball, contradicting strict convexity there. The global minimizer is unique and satisfies $\|\Delta_\pi\|_2 < \rho$. Moreover, $\|E_\pi\|_2 + L_{\mathrm{tr}}\rho/2 \le \mu/4 + \mu/8 = 3\mu/8 < \mu/2$, so Eq. A.65 holds.

Relation to the main stability event For the fixed design π, write $H := H_{\mathrm{tr}}(\hat{\theta})$ within this comparison. Assume the selected minimizers and segment Hessians are well defined on all supported draws, as required in Corollary A.1. On $\mathcal{G}_\pi$, integration of the Hessian Lipschitz bound along the retraining segment gives

$$\|\overline{H}_\pi - H\|_2 \le \|E_\pi\|_2 + \frac{L_{\mathrm{tr}}}{2}\|\Delta_\pi\|_2 \le \frac{3\mu}{8} < \frac{1}{2}\lambda_{\min}(H).$$

The last inequality uses $\lambda_{\min}(H) \ge \mu > 0$. Consequently,

$$\mathcal{G}_\pi \subseteq \mathcal{A}_\pi, \qquad \mathbb{P}_W(\mathcal{A}_\pi^c) \le \mathbb{P}_W(\mathcal{G}_\pi^c) = O(B^{-2}). \tag{A.81}$$

The fourth moment on the larger event requires a separate bound. Sampled stationarity and the lower bound $\overline{H}_\pi \succeq (\mu/2)I$ give $\|\Delta_\pi\|_2 \le 2\|g_\pi\|_2/\mu$ on $\mathcal{G}_\pi$. Thus

$$\mathbb{E}_W\left[\|\Delta_\pi\|_2^4 \mathbf{1}_{\mathcal{G}_\pi}\right] \le \left(\frac{2}{\mu}\right)^4 \mathbb{E}_W\|g_\pi\|_2^4 = O(1)O(B^{-2}) = O(B^{-2}).$$

Using $(a+b)^4 \le 8(a^4+b^4)$ in the integrated Hessian bound, and applying Eqs. A.59 and A.60, yields

$$\begin{aligned}
&\mathbb{E}_W\left[\|\overline{H}_\pi - H\|_2^4 \mathbf{1}_{\mathcal{G}_\pi}\right] \\
&\quad \le 8\mathbb{E}_W\|E_\pi\|_2^4 + \frac{L_{\mathrm{tr}}^4}{2}\mathbb{E}_W\left[\|\Delta_\pi\|_2^4 \mathbf{1}_{\mathcal{G}_\pi}\right] \\
&\quad = O(B^{-2}) + O(1)O(B^{-2}) = O(B^{-2}).
\end{aligned}$$

The per-example Hessian bound used in the perturbation calculation also gives

$$\lambda_{\min}(H) \le \|H\|_2 \le \frac{1}{n}\sum_i \|\nabla_\theta^2 \ell(Z_i;\hat{\theta})\|_2 \le \frac{1}{n}\sum_i \|\nabla_\theta^2 \ell(Z_i;\hat{\theta})\|_F \le K.$$

On $\mathcal{A}_\pi \setminus \mathcal{G}_\pi$, the definition of $\mathcal{A}_\pi$ therefore bounds $\|\overline{H}_\pi - H\|_2$ by $K/2$, so

$$\begin{aligned}
&\mathbb{E}_W\left[\|\overline{H}_\pi - H\|_2^4 \mathbf{1}_{\mathcal{A}_\pi \setminus \mathcal{G}_\pi}\right] \\
&\quad \le \frac{K^4}{16}\mathbb{P}_W(\mathcal{G}_\pi^c) = O(1)O(B^{-2}) = O(B^{-2}).
\end{aligned}$$

Adding the two disjoint contributions gives the $O(B^{-2}) + O(B^{-2}) = O(B^{-2})$ fourth-moment bound on $\mathcal{A}_\pi$ in Eq. A.6. For the exceptional displacement moments, set inclusion gives

$$\mathbb{E}_W\left[\|\Delta_\pi\|_2^k \mathbf{1}_{\mathcal{A}_\pi^c}\right] \le \mathbb{E}_W\left[\|\Delta_\pi\|_2^k \mathbf{1}_{\mathcal{G}_\pi^c}\right], \qquad k = 2, 3.$$

If the right-hand sides satisfy the auxiliary event's tail conditions with $q = B^{-1/2}$, their respective orders are $O(q^4) = O(B^{-2})$ and $O(q^3) = O(B^{-3/2})$. The same bounds then hold on $\mathcal{A}_\pi^c$.

Frozen-feature softmax With fixed features, logits are affine in the trainable head parameters and the token cross-entropy is convex and nonquadratic. For bounded features and a fixed head dimension, its gradients and Hessians are bounded and its Hessian is Lipschitz. The preceding argument applies when the common-logit shift has been removed, for example by fixing a reference class, and the full-data Hessian has a minimum eigenvalue bounded below. These curvature and identifiability requirements are additional conditions, not consequences of feature boundedness alone.

This argument establishes localization on $\mathcal{G}_\pi$. To obtain the proposition's expectation guarantee, a softmax instantiation must also have finite selected minimizers on exceptional draws and satisfy Eq. A.67. Unregularized softmax can have separable sampled subsets with no finite minimizer; feature boundedness does not resolve this issue. If selected displacements are uniformly bounded, the $O(B^{-2})$ failure probability established above supplies the exceptional moments by the bound following Proposition 8's proof. For nonconvex training losses, the local argument alone cannot identify the selected global minimizer, so localization remains an explicit assumption.

A complete sufficient condition for regularized convex losses The local argument above separates localization from the existence and exceptional moments of the training outputs. The following condition supplies all of them without specifying a particular loss formula. Replace the training objectives by

$$F^{\lambda}_{\mathrm{tr}}(\theta) := \frac{1}{n}\sum_i \ell(Z_i;\theta) + \frac{\lambda}{2}\|\theta\|_2^2, \qquad F^{\lambda}_{\mathrm{tr},\pi}(\theta;W) := \frac{1}{n}\sum_i \frac{W_i}{\pi_i}\ell(Z_i;\theta) + \frac{\lambda}{2}\|\theta\|_2^2, \tag{A.82}$$

where $\lambda > 0$ is fixed independently of B. The penalty is deterministic and is not multiplied by HT weights. In the following regularized result and the two model examples in Appendices A.4.4 and A.4.5, set $\theta_0 = \hat{\theta}$; $\hat{\theta}$ and $\hat{\theta}_\pi(W)$ denote the respective regularized minimizers, and $H_{\mathrm{tr}}(\hat{\theta}) := \nabla^2_\theta F^{\lambda}_{\mathrm{tr}}(\hat{\theta})$. On the empty draw, use the regularized minimizer 0, rather than the baseline convention for the unregularized objective. Keep the validation loss unregularized, and write $\mathcal{R}_\lambda$ and $\widetilde{\mathcal{R}}_\lambda$ for the resulting actual and surrogate risks. In the surrogate, use $\mathcal{I}_{\hat{\theta}}(Z_i) = -H_{\mathrm{tr}}(\hat{\theta})^{-1}\nabla_\theta \ell(Z_i;\hat{\theta})$ to define a_i.

Proposition 9 (Regularity for convex smooth losses with fixed ridge)**.** *Suppose every training and validation data loss is nonnegative, convex, and twice continuously differentiable on the parameter space. For every parameter and example, suppose its gradient norm is at most G, its Hessian Frobenius norm is at most K, and its Hessian is L_ℓ-Lipschitz in operator norm. Let G, K, L_ℓ be finite and independent of B. Fix $0 < c < 1$ and $\lambda > 0$. For every design sequence in $\mathcal{P}_\alpha(B)$ with $\alpha = cB/n$, the exact-stationarity, smoothness, stability, and moment conditions of Corollary A.1 hold for these regularized objectives, with constants depending only on c, λ, G, K, L_ℓ. The actual-risk minimum is attained. For positive exact scores, the risk comparison in Eq. A.40 holds for $\mathcal{R}_\lambda$.*

Proof. Existence, curvature, and displacement. Every data loss is nonnegative and has positive semidefinite Hessian. Each objective in Eq. A.82 is continuous, bounded below by $\lambda\|\theta\|_2^2/2$, and has Hessian at least λI. Coercivity gives a finite minimizer, and strict convexity makes it unique for every draw. Write $H := H_{\mathrm{tr}}(\hat{\theta})$; the derivative bounds give $\lambda I \preceq H \preceq (\lambda + K)I$. With $s_i := \nabla_\theta \ell(Z_i;\hat{\theta})$ and $H_i := \nabla^2_\theta \ell(Z_i;\hat{\theta})$, full-data stationarity and cancellation of the deterministic penalty give

$$\begin{aligned} g_\pi &:= \nabla_\theta F^{\lambda}_{\mathrm{tr},\pi}(\hat{\theta};W) = \frac{1}{n}\sum_i \left(\frac{W_i}{\pi_i} - 1\right) s_i, \\ E_\pi &:= \nabla^2_\theta F^{\lambda}_{\mathrm{tr},\pi}(\hat{\theta};W) - H = \frac{1}{n}\sum_i \left(\frac{W_i}{\pi_i} - 1\right) H_i. \end{aligned} \tag{A.83}$$

In particular, $\Delta^{\mathrm{lin}}_\pi = -H^{-1}g_\pi$ has zero mean. Global strong convexity and sampled stationarity give

$$\begin{aligned} \lambda\|\Delta_\pi\|_2^2 &\le \Delta_\pi^\top[\nabla_\theta F^{\lambda}_{\mathrm{tr},\pi}(\hat{\theta}+\Delta_\pi;W) - \nabla_\theta F^{\lambda}_{\mathrm{tr},\pi}(\hat{\theta};W)] \\ &= -\Delta_\pi^\top g_\pi \le \|\Delta_\pi\|_2\|g_\pi\|_2. \end{aligned}$$

If $\Delta_\pi \neq 0$, division by its norm proves the following bound; it also holds when $\Delta_\pi = 0$:

$$\|\Delta_\pi\|_2 \le \lambda^{-1}\|g_\pi\|_2 \quad \text{for every draw.} \tag{A.84}$$

The spectral event and fourth moments. The centered sums in Eq. A.83 have exactly the form treated in the perturbation calculation above. The probability floor gives

$$\mathbb{E}_W\|g_\pi\|_2^4 = O(B^{-2}), \qquad \mathbb{E}_W\|E_\pi\|_2^4 = O(B^{-2}), \qquad \mathbb{E}_W\left|\frac{1}{n}\sum_i \frac{W_i}{\pi_i} - 1\right|^4 = O(B^{-2}),$$

with constants independent of the particular design. Within this proof, define the auxiliary event

$$\mathcal{G}_\pi := \left\{\|E_\pi\|_2 \le \frac{\lambda}{4}, \quad \|g_\pi\|_2 \le \frac{\lambda^2}{4(1+L_\ell)}, \quad \frac{1}{n}\sum_i \frac{W_i}{\pi_i} \le 2\right\}.$$

Markov's inequality at the three fixed thresholds and a union bound give

$$\mathbb{P}_W(\mathcal{G}_\pi^c) \le \left(\frac{4}{\lambda}\right)^4 \mathbb{E}_W\|E_\pi\|_2^4 + \left(\frac{4(1+L_\ell)}{\lambda^2}\right)^4 \mathbb{E}_W\|g_\pi\|_2^4 + \mathbb{E}_W\left|\frac{1}{n}\sum_i \frac{W_i}{\pi_i} - 1\right|^4$$
$$= O(B^{-2}) + O(B^{-2}) + O(B^{-2}) = O(B^{-2}).$$

The regularization Hessian is constant. On $\mathcal{G}_\pi$, the sampled loss Hessian has Lipschitz constant at most $2L_\ell$. Define $\overline{H}_\pi$ by Eq. A.4 for the regularized objective. Integration along the retraining segment gives

$$\|\overline{H}_\pi - H\|_2 \le \|E_\pi\|_2 + \int_0^1 2L_\ell t\|\Delta_\pi\|_2\,\mathrm{d}t$$
$$\le \|E_\pi\|_2 + \frac{L_\ell}{\lambda}\|g_\pi\|_2 \le \frac{\lambda}{4} + \frac{\lambda}{4} = \frac{\lambda}{2} \le \frac{1}{2}\lambda_{\min}(H).$$

Thus $\mathcal{G}_\pi \subseteq \mathcal{A}_\pi$, where $\mathcal{A}_\pi$ is precisely the spectral event in Eq. A.5 with the regularized Hessians. It follows that $\mathbb{P}_W(\mathcal{A}_\pi^c) \le \mathbb{P}_W(\mathcal{G}_\pi^c) = O(B^{-2})$. For the required fourth moment on the full event, first use $(x+y)^4 \le 8(x^4+y^4)$ on $\mathcal{G}_\pi$:

$$\mathbb{E}_W[\|\overline{H}_\pi - H\|_2^4 \mathbf{1}_{\mathcal{G}_\pi}] \le 8\mathbb{E}_W\|E_\pi\|_2^4 + 8\frac{L_\ell^4}{\lambda^4}\mathbb{E}_W\|g_\pi\|_2^4$$
$$= O(B^{-2}) + O(1)O(B^{-2}) = O(B^{-2}).$$

On $\mathcal{A}_\pi \setminus \mathcal{G}_\pi$, the spectral event itself bounds $\|\overline{H}_\pi - H\|_2$ by $\lambda_{\min}(H)/2 \le (\lambda+K)/2$. Hence

$$\mathbb{E}_W[\|\overline{H}_\pi - H\|_2^4 \mathbf{1}_{\mathcal{A}_\pi \setminus \mathcal{G}_\pi}] \le \left(\frac{\lambda+K}{2}\right)^4 \mathbb{P}_W(\mathcal{G}_\pi^c) = O(1)O(B^{-2}) = O(B^{-2}).$$

Adding these disjoint contributions proves Eq. A.6.

Exceptional moments and risk transfer. The global displacement bound and Hölder's inequality give

$$\mathbb{E}_W[\|\Delta_\pi\|_2^2 \mathbf{1}_{\mathcal{A}_\pi^c}] \le \lambda^{-2}(\mathbb{E}_W\|g_\pi\|_2^4)^{1/2}\mathbb{P}_W(\mathcal{A}_\pi^c)^{1/2} = O(B^{-1})O(B^{-1}) = O(B^{-2}),$$
$$\mathbb{E}_W[\|\Delta_\pi\|_2^3 \mathbf{1}_{\mathcal{A}_\pi^c}] \le \lambda^{-3}(\mathbb{E}_W\|g_\pi\|_2^4)^{3/4}\mathbb{P}_W(\mathcal{A}_\pi^c)^{1/4} = O(B^{-3/2})O(B^{-1/2}) = O(B^{-2}).$$

The latter is also $O(B^{-3/2})$ for $B \ge 1$, as required. Validation Hessians have global Lipschitz constant L_ℓ and norm at most K. This verifies the stability and exact-stationarity conditions in Corollary A.1 for this regularized setting. The underlying accuracy argument uses $H\Delta_\pi^{\mathrm{lin}} = -g_\pi$, $g_\pi + \overline{H}_\pi\Delta_\pi = 0$, and the validation Taylor expansion. The first identity follows from Eq. A.83; the second follows by integrating the sampled regularized gradient, which is zero at its minimizer, including on the empty draw. Thus the same risk bounds follow for $\mathcal{R}_\lambda$ and $\widetilde{\mathcal{R}}_\lambda$.

Existence of the competing optimum. For fixed data and B, the feasible set $\mathcal{P}_\alpha(B)$ is nonempty and compact. To check continuity of the actual risk, fix a draw w and two designs π, π'. Strong convexity, the two stationarity equations, and the uniform gradient bound give

$$\lambda\|\hat{\theta}_\pi(w) - \hat{\theta}_{\pi'}(w)\|_2 \le \|\nabla_\theta F^\lambda_{\mathrm{tr},\pi}(\hat{\theta}_{\pi'}(w); w)\|_2$$
$$= \left\|\frac{1}{n}\sum_i w_i\left(\frac{1}{\pi_i} - \frac{1}{\pi_i'}\right)\nabla_\theta \ell(Z_i; \hat{\theta}_{\pi'}(w))\right\|_2$$
$$\le \frac{G}{n}\sum_i w_i\left|\frac{1}{\pi_i} - \frac{1}{\pi_i'}\right|.$$

The first line follows by applying the monotonicity inequality used for Eq. A.84 to the two parameters in the objective indexed by π. The right-hand side tends to zero as $\pi' \to \pi$, since all coordinates are at least $\alpha > 0$. Each fitted parameter is therefore continuous in the design. Express $\mathcal{R}_\lambda(\pi)$ as the finite sum over $w \in \{0,1\}^n$ of $F_{\mathrm{val}}(\hat{\theta}_\pi(w))$ times the Bernoulli product probability. Both factors are continuous, so the risk is continuous and attains a minimum on the compact feasible set.

All constants above depend only on c, λ, G, K, L_ℓ; the bounds apply separately to the proposed and minimizing designs, each under its own sampling law. Moreover, $0 \le a_i \le KG^2/\lambda^2 = O(1)$. For positive coefficients, the regularity bounds above and surrogate optimality give the comparison in Eq. A.40. If the validation gradient norm is also $o(1)$, the $o(B^{-1})$ conclusion of Theorem 1 applies. Under the additional relative accuracy condition, the score bound gives Corollary A.3. □

Combining regularity with the gradient condition The proof uses the derivative bounds only through finite sums and products of G, K, L_ℓ and inverse powers of the fixed positive c, λ. Thus if these derivative bounds depend on random data and are $O_p(1)$, while convexity and smoothness hold on data events of probability tending to one, every normalized moment and tail control in Eq. A.49 is $O_p(1)$. For example, the conditional gradient fourth moment is bounded by $G^4[1/(c^3B^3) + 3/(c^2B^2)]$, so its product with B^2 is $O_p(1)(B^{-1}+1) = O_p(1)$. The conditional Hessian fourth moment has the same bound with K in place of G. The auxiliary-event bound gives

$$B^2\mathbb{P}_W(\mathcal{G}_\pi^c) \le O_p(1) + (1+O_p(1))^4 O_p(1) + O(1) = O_p(1).$$

Since $\mathcal{G}_\pi \subseteq \mathcal{A}_\pi$, this also controls $B^2\mathbb{P}_W(\mathcal{A}_\pi^c)$. The two parts of the averaged-Hessian fourth moment are products of tight derivative bounds and these tight normalized moments, so their sum after multiplication by B^2 is $O_p(1)$. The exceptional second-moment inequality gives

$$B^2\mathbb{E}_W[\|\Delta_\pi\|_2^2 \mathbf{1}_{\mathcal{A}_\pi^c}] \le \lambda^{-2}(B^2\mathbb{E}_W\|g_\pi\|_2^4)^{1/2}[B^2\mathbb{P}_W(\mathcal{A}_\pi^c)]^{1/2} = O_p(1)O_p(1) = O_p(1).$$

For the exceptional third moment, the corresponding product is $O_p(B^{-2})^{3/4}O_p(B^{-2})^{1/4} = O_p(B^{-2})$; its product with $B^{3/2}$ is $O_p(B^{-1/2})$, hence $O_p(1)$. The validation Hessian and its Lipschitz constant are bounded by K and L_ℓ, and the optimization residuals vanish. These implications apply uniformly to the design class, including both competing designs.

When the loss bounds instead hold uniformly over the support of the data distribution, the constants are deterministic for every dataset. For independent samples with bounded total score variances, the strong-convexity argument in Appendix A.4.2 supplies the root-n training rate. If the same regularized population training point has $\delta_{\mathrm{tv}} \to 0$, that subsection supplies a vanishing validation gradient. Subject to positive scores, these facts jointly verify Proposition 6. The shared population point must include the chosen penalty; stability alone places no restriction on the training–validation mismatch.

A.4.4 Regularity verification for regularized frozen-feature softmax

We verify the loss conditions of Proposition 9 for frozen-feature softmax, using the regularized objectives and notation in Eq. A.82. For a frozen-feature head, each token loss takes the form

$$\log\left[\sum_{k=1}^{J}\exp(\theta^\top\phi_{itk})\right] - \theta^\top\phi_{it,y_{it}}, \qquad \|\phi_{itk}\|_2 \le R_\phi, \tag{A.85}$$

where t indexes supervised tokens, y_{it} is the observed label, and ϕ_{itk} encodes the frozen features for class k in the head parameter vector. The per-example loss is the average over its tokens. Assume the same feature bound for validation, with $R_\phi > 0$ fixed.

Proposition 10 (Structural conditions covering the competing design)**.** *In the regularized softmax setting above, fix $0 < c < 1$ and $\alpha = cB/n$. For every sequence of designs in $\mathcal{P}_\alpha(B)$, the stationarity, smoothness, stability, and moment conditions of Corollary A.1 hold for the regularized objectives, with constants depending only on c, λ, R_ϕ. The actual-risk minimum on this class is attained. Consequently, the risk accuracy and comparison results apply to $\mathcal{R}_\lambda$ and $\widetilde{\mathcal{R}}_\lambda$. For positive exact scores, Eq. A.40 applies; the $o(B^{-1})$ conclusion of Theorem 1 follows if additionally $\|\nabla_\theta F_{\mathrm{val}}(\hat\theta)\|_2 = o(1)$. For positive exact scores, Corollary A.3 in Appendix A.5 applies to approximate scores satisfying Eq. A.92.*

Proof. For one token, suppress i, t and set $p_k := \exp(\theta^\top\phi_k)/\sum_j \exp(\theta^\top\phi_j)$ and $\bar\phi := \sum_k p_k\phi_k$. Direct differentiation gives

$$\nabla\ell = \bar\phi - \phi_y, \qquad \nabla^2\ell = \sum_k p_k(\phi_k-\bar\phi)(\phi_k-\bar\phi)^\top, \qquad Dp_k[u] = p_k u^\top(\phi_k-\bar\phi).$$

Here D denotes a directional derivative. Since $\|\bar\phi\|_2 \le R_\phi$ and $\|vv^\top\|_F = \|v\|_2^2$, these formulas give $\|\nabla\ell\|_2 \le 2R_\phi$ and $\|\nabla^2\ell\|_F \le 4R_\phi^2$. Differentiating the covariance in directions u, v, w yields

$$D^3\ell[u,v,w] = \sum_k p_k[u^\top(\phi_k-\bar\phi)][v^\top(\phi_k-\bar\phi)][w^\top(\phi_k-\bar\phi)].$$

The terms differentiating $\bar{\phi}$ vanish because $\sum_k p_k(\phi_k - \bar{\phi}) = 0$; the displayed terms come from $Dp_k[w]$. Each centered feature has norm at most $2R_\phi$, so $|D^3\ell[u,v,w]| \le 8R_\phi^3\|u\|_2\|v\|_2\|w\|_2$. Integrating this derivative between any two parameter vectors and taking the supremum over unit u, v proves an operator-norm Hessian Lipschitz bound. Averaging over tokens preserves all three bounds. We may thus take

$$G := 2R_\phi, \qquad K := 4R_\phi^2, \qquad L_\ell := 8R_\phi^3$$

for training and validation, uniformly in the parameter.

Each token loss is nonnegative because the log-sum-exp includes the observed-label logit. Its Hessian is a covariance matrix and is positive semidefinite, so the loss is convex. The displayed derivative bounds and Proposition 9 verify the claimed conditions, with constants depending only on c, λ, R_ϕ. In addition, $0 \le a_i \le KG^2/\lambda^2$, because the validation Hessian is positive semidefinite and bounded by K, and $\|\mathcal{I}_{\hat{\theta}}(Z_i)\|_2 \le G/\lambda$. This coefficient bound also verifies the bounded-score requirement of Corollary A.3 when its relative score-accuracy condition holds. □

These regularity constants are uniform over data realizations with the stated feature bound. Positive scores and a small validation gradient are additional properties; the following example verifies them jointly with regularity under random data.

A.4.5 A BINARY CROSS-ENTROPY EXAMPLE UNDER RANDOM DATA

We first retain an explicit calculation with fixed empirical proportions, then use the corresponding population probabilities to verify joint conditions over random datasets.

Fixed empirical proportions The following proportions give a zero validation gradient and nonuniform surrogate-optimal sampling. Consider one token per example and one scalar parameter, with

$$\ell[(x,y);\theta] = \log(1+e^{x\theta}) - yx\theta, \qquad \sigma(u) := (1+e^{-u})^{-1}, \qquad \lambda := \frac{1}{6\log 2}.$$

Let n be a multiple of 20. Half of the training examples have $x = 1$, with positive-label fraction $4/5$; the other half have $x = 2$, with positive-label fraction $9/10$. At $\hat{\theta} = \log 2$, the predicted positive probabilities are $\sigma(\hat{\theta}) = 2/3$ and $\sigma(2\hat{\theta}) = 4/5$. The regularized training gradient is

$$\nabla_\theta F_{\mathrm{tr}}^\lambda(\hat{\theta}) = \frac{1}{2}\left(\frac{2}{3} - \frac{4}{5}\right) + \frac{1}{2}\,2\left(\frac{4}{5} - \frac{9}{10}\right) + \lambda\log 2 = -\frac{1}{6} + \frac{1}{6} = 0.$$

Strict convexity identifies $\log 2$ as the unique full-data fit.

For validation, use equally sized groups with $x = 1$ and $x = 2$, but positive-label fractions $2/3$ and $4/5$, respectively; a validation size divisible by 30 realizes these proportions. The validation gradient vanishes within each feature group. Using $\nabla_\theta^2\ell = x^2\sigma(x\theta)[1-\sigma(x\theta)]$ gives

$$H_{\mathrm{val}}(\hat{\theta}) = \frac{1}{2}\frac{2}{9} + \frac{1}{2}\frac{16}{25} = \frac{97}{225}, \qquad H_{\mathrm{tr}}(\hat{\theta}) = \frac{97}{225} + \lambda.$$

For the four training types $(x,y) = (1,1), (1,0), (2,1), (2,0)$, the gradients and exact scores are

$$s_i \in \left\{-\frac{1}{3}, \frac{2}{3}, -\frac{2}{5}, \frac{8}{5}\right\}, \qquad a_i = \frac{97/225}{(97/225+\lambda)^2}s_i^2 > 0. \tag{A.86}$$

Their training frequencies are $2/5, 1/10, 9/20, 1/20$, giving $n^{-1}\sum_i |s_i| = 23/50$ and $n^{-1}\sum_i s_i^2 = 13/45$. Take $B = n/4$ and $c = 1/2$, so the probability floor is $1/8$. The surrogate-optimal design is

$$\pi^\star_{\alpha,i} = \frac{B}{n}\frac{|s_i|}{23/50}.$$

Its smallest and largest entries are $25/138 > 1/8$ and $20/23 < 1$, so neither constraint changes these probabilities. They are nonuniform, with ratios $1/3 : 2/3 : 2/5 : 8/5$ across the four types. Direct substitution into the surrogate gives

$$\widetilde{\mathcal{R}}_\lambda(\pi^\star_\alpha) - F_{\mathrm{val}}(\hat{\theta}) = \frac{97/225}{2B(97/225+\lambda)^2}\left[\left(\frac{23}{50}\right)^2 - \frac{1}{4}\frac{13}{45}\right] = \Theta(B^{-1}), \tag{A.87}$$

because the bracket is a strictly positive constant. Proposition 10 covers both the surrogate-optimal design and the actual-risk optimizer. The zero validation gradient removes the B^{-1} bias contribution in Eq. A.40, giving

$$0 \le \mathcal{R}_\lambda(\pi_\alpha^\star) - \mathcal{R}_\lambda(\pi_{\mathrm{opt}}) = O(B^{-3/2}).$$

Uniform also satisfies the stability conditions by Proposition 10. The empirical variance of the square-root scores is the positive constant

$$\begin{aligned}\mathrm{Var}_n(\sqrt{a_i}) &= \frac{97/225}{(97/225+\lambda)^2}\left[\frac{13}{45} - \left(\frac{23}{50}\right)^2\right] \\ &= \frac{1739}{22500}\frac{97/225}{(97/225+\lambda)^2} > 0.\end{aligned}$$

The interior probabilities verified above and the zero validation gradient therefore satisfy all conditions of Proposition 5. In particular, Eq. A.46 gives

$$\mathcal{R}_\lambda(\pi^{\mathrm{unif}}) - \mathcal{R}_\lambda(\pi_{\mathrm{opt}}) = \frac{1739}{45000B}\frac{97/225}{(97/225+\lambda)^2} + O(B^{-3/2}) = \Theta(B^{-1}).$$

For approximate scores satisfying Eq. A.92 with $\eta_B = O(B^{-1/4})$, Corollary A.3 adds a term $O(\eta_B^2/B) = O(B^{-1/2})B^{-1} = O(B^{-3/2})$, giving the same regret rate. Relative to the increment in Eq. A.87, the surrogate-optimal rule's regret is $O(B^{-3/2})/\Theta(B^{-1}) = O(B^{-1/2}) \to 0$. This example realizes the sufficient conditions with a nonquadratic cross-entropy loss and a nonzero fitted parameter.

Random training and validation samples Now draw the training and validation samples independently, each with independent and identically distributed observations. Both populations assign probability $1/2$ to each of $x = 1$ and $x = 2$, and their label probabilities are

	$x=1$	$x=2$
$\mathbb{P}_{\mathrm{tr}}(y=1 \mid x)$	$4/5$	$9/10$
$\mathbb{P}_{\mathrm{val}}(y=1 \mid x)$	$2/3$	$4/5.$

(A.88)

Here $\mathbb{P}_{\mathrm{tr}}$ and $\mathbb{P}_{\mathrm{val}}$ are the single-example population laws corresponding to $\mathbb{E}_{\mathrm{tr}}$ and $\mathbb{E}_{\mathrm{val}}$. Keep $\lambda = 1/(6\log 2)$. The sample sizes need not realize any exact proportions or satisfy divisibility restrictions. The fit $\hat{\theta}$, scores a_i, and probabilities $\pi_\alpha^\star$ are recomputed from each realized dataset; $\log 2$ is now a population reference, not an identity for the empirical fit.

Proposition 11 (Joint verification in the random binary model)**.** *For the populations in Eq. A.88, let $n, m \ge 1$, $B \to \infty$, $B \le n$, and $\alpha = cB/n$ with fixed $0 < c < 1$. The regularity conditions hold with deterministic constants uniform over all datasets and designs, and $a_i > 0$ for every training example. The population training point is $\theta_{\mathrm{tr}}^\circ = \log 2$, with $\delta_{\mathrm{tv}} = 0$, and*

$$\mathbb{E}_{\mathrm{data}}|\hat{\theta} - \log 2|^2 \le \frac{47}{180\lambda^2 n}, \tag{A.89}$$

$$\mathbb{E}_{\mathrm{data}}|\nabla F_{\mathrm{val}}(\hat{\theta})|^2 \le \frac{47}{90\lambda^2 n} + \frac{194}{225m}. \tag{A.90}$$

Consequently the validation gradient has order $O_p(n^{-1/2} + m^{-1/2})$ and

$$0 \le \mathcal{R}_\lambda(\pi_\alpha^\star) - \mathcal{R}_\lambda(\pi_{\mathrm{opt}}) = O_p\left(\frac{1}{B\sqrt{n}} + \frac{1}{B\sqrt{m}} + B^{-3/2}\right). \tag{A.91}$$

The excess risk is $o_p(B^{-1})$ when $n, m \to \infty$, and $O_p(B^{-3/2})$ when $m \gtrsim B$.

Proof. At $\theta = \log 2$, the training scores take the four values in Eq. A.86 with the corresponding population probabilities. Their mean, second moment, and variance are

$$\mathbb{E}_{\mathrm{tr}}\nabla\ell(Z;\log 2) = -\frac{1}{6}, \qquad \mathbb{E}_{\mathrm{tr}}|\nabla\ell(Z;\log 2)|^2 = \frac{13}{45}, \qquad \frac{13}{45} - \frac{1}{36} = \frac{47}{180}.$$

The penalty contributes $\lambda \log 2 = 1/6$, so the regularized population training gradient is zero. Its Hessian is at least λ, identifying the unique population minimizer. For validation, the conditional

label mean equals $\sigma(x \log 2)$ for both feature values. Thus $\mathbb{E}_{\mathrm{val}} \nabla \ell(Z; \log 2) = 0$, giving $\delta_{\mathrm{tv}} = 0$, and the score variance is

$$\mathbb{E}_{\mathrm{val}} |\nabla \ell(Z; \log 2)|^2 = \frac{1}{2}\frac{2}{9} + \frac{1}{2}\frac{16}{25} = \frac{97}{225}.$$

For every training dataset, strong convexity and empirical stationarity give

$$|\hat{\theta} - \log 2| \le \lambda^{-1} |\nabla F_{\mathrm{tr}}^{\lambda}(\log 2)|.$$

The latter gradient is an average of centered independent scores, so its second moment equals $47/(180n)$. Squaring the displacement inequality and taking data expectation proves Eq. A.89. Every empirical validation Hessian obeys

$$0 < H_{\mathrm{val}}(\theta) = \frac{1}{m} \sum_j (x_j^{\mathrm{val}})^2 \sigma(x_j^{\mathrm{val}} \theta)[1 - \sigma(x_j^{\mathrm{val}} \theta)] \le 1$$

for finite θ, since $x_j^{\mathrm{val}} \in \{1, 2\}$ and $0 < \sigma(u)[1 - \sigma(u)] \le 1/4$. Integrating the Hessian bounds the gradient difference by $|\hat{\theta} - \log 2|$. The inequality $(a + b)^2 \le 2a^2 + 2b^2$ gives

$$\begin{aligned} \mathbb{E}_{\mathrm{data}} |\nabla F_{\mathrm{val}}(\hat{\theta})|^2 &\le 2\mathbb{E}_{\mathrm{data}} |\nabla F_{\mathrm{val}}(\log 2)|^2 + 2\mathbb{E}_{\mathrm{data}} |\hat{\theta} - \log 2|^2 \\ &\le \frac{194}{225m} + \frac{47}{90\lambda^2 n}. \end{aligned}$$

This proves Eq. A.90. Markov's inequality at threshold $M(n^{-1/2} + m^{-1/2})$ bounds the gradient-norm exceedance probability by a constant times M^{-2}, proving the stated O_p rate. At any fixed positive threshold, the same bound tends to zero when $n, m \to \infty$.

The feature bound $|x| \le 2$ and fixed ridge allow Proposition 10 to verify regularity for every dataset and every design. Every fitted parameter is finite. For each training example, $x_i[\sigma(x_i \hat{\theta}) - y_i] \ne 0$ because $x_i \ne 0$, $y_i \in \{0, 1\}$, and $0 < \sigma(x_i \hat{\theta}) < 1$. The positive scalar validation Hessian and positive training Hessian therefore give

$$a_i = \frac{H_{\mathrm{val}}(\hat{\theta})}{H_{\mathrm{tr}}(\hat{\theta})^2} |\nabla \ell(Z_i; \hat{\theta})|^2 > 0.$$

The regularized fits are continuous in the data and design, and the conditional risk is continuous in the design on its compact feasible set. A measurable minimizing design can consequently be selected; in this finite-support data model it suffices to fix one minimizer for each of the finitely many datasets at each pair of sample sizes. All hypotheses of Proposition 6 now hold. Substitution of the gradient rate gives $O_p(n^{-1/2})B^{-1} + O_p(m^{-1/2})B^{-1} + O_p(B^{-3/2})$, which is Eq. A.91. Multiplying by B gives terms of orders $O_p(n^{-1/2})$, $O_p(m^{-1/2})$, and $O_p(B^{-1/2})$, all tending to zero when both sample sizes grow. When $m \gtrsim B$ and $n \ge B$, each term in the unscaled bound is $O_p(B^{-3/2})$. □

The exact probability values and the explicit uniform-sampling comparison above concern the fixed empirical proportions. The random model establishes joint regularity and vanishing-gradient conditions without requiring those proportions to hold exactly. Its probability rates refer to the conditional expected risks; they do not assert a fixed multiple of $B^{-3/2}$ with probability tending to one merely from n, m being proportional to B. Strict separation from scalar IF sampling with nonzero coefficients is established by the two-dimensional example in Appendix A.3.

A.5 Sensitivity to score and optimization errors

In practice, the exact coefficients a_i are replaced by the sketched estimates $\widehat{a}_i$ of Eq. 22. Since $\widehat{a}_i$ may vanish numerically, we use the effective coefficients $\widetilde{a}_i := \max\{\widehat{a}_i, \tau\}$ with a small $\tau > 0$, and let $\widehat{\pi}$ be the rule in Eq. 14 with a_i replaced by $\widetilde{a}_i$, that is, the exact minimizer of $\sum_i \widetilde{a}_i / \pi_i$ over $\mathcal{P}_\alpha(B)$. Both a_i and $\widetilde{a}_i$ are evaluated at the same checkpoint θ_0. We measure their total relative error by

$$|\widetilde{a}_i - a_i| \le \eta_B a_i \quad \text{for all } i, \qquad 0 \le \eta_B \le \eta_0 < 1. \tag{A.92}$$

This condition covers the combined effect of the GNH curvature approximation, the sketch, the damping ζ, and the floor τ. A norm-preservation guarantee for the sketch alone does not establish it, and we do not model errors in calibrating κ.

Proposition 12 (Quadratic effect of relative score error)**.** *Let $a_i > 0$, and let $\widehat{\pi}$ minimize $\sum_i \widetilde{a}_i/\pi_i$ over $\mathcal{P}_\alpha(B)$, where $|\widetilde{a}_i - a_i| \le \eta_B a_i$ and $0 \le \eta_B < 1$. For any $0 < \alpha < 1$ with $\alpha n \le B \le n$,*

$$\begin{aligned} 0 &\le \widetilde{\mathcal{R}}(\widehat{\pi}) - \widetilde{\mathcal{R}}(\pi^\star_\alpha) \\ &\le \frac{\eta_B^2}{1-\eta_B^2}\frac{1}{2n^2}\sum_i \frac{a_i}{\pi^\star_{\alpha,i}} \le \frac{a_{\max}}{2B}\frac{\eta_B^2}{1-\eta_B^2}, \qquad a_{\max} := \max_i a_i. \end{aligned} \tag{A.93}$$

The bound permits both lower and upper probability constraints to be active.

Proof. In this proof write $S(\pi) := \sum_i a_i/\pi_i$ and $\mathcal{E} := S(\widehat{\pi}) - S(\pi^\star_\alpha) \ge 0$. For positive x, y, the exact identity

$$\frac{1}{y} - \frac{1}{x} = -\frac{y-x}{x^2} + \frac{(y-x)^2}{yx^2}$$

gives

$$\mathcal{E} = \nabla S(\pi^\star_\alpha)^\top(\widehat{\pi} - \pi^\star_\alpha) + D, \qquad D := \sum_i \frac{a_i(\widehat{\pi}_i - \pi^\star_{\alpha,i})^2}{\widehat{\pi}_i(\pi^\star_{\alpha,i})^2}. \tag{A.94}$$

The feasible set is convex, so $\pi^\star_\alpha + t(\widehat{\pi} - \pi^\star_\alpha)$ is feasible for $0 \le t \le 1$. Its objective is minimized at $t = 0$; dividing its nonnegative increment by $t > 0$ and taking $t \downarrow 0$ proves $\nabla S(\pi^\star_\alpha)^\top(\widehat{\pi} - \pi^\star_\alpha) \ge 0$. Consequently $D \le \mathcal{E}$, including at clipped optima.

Optimality for the approximate scores gives

$$\sum_i \frac{\widetilde{a}_i}{\widehat{\pi}_i} \le \sum_i \frac{\widetilde{a}_i}{\pi^\star_{\alpha,i}}.$$

Subtract the corresponding terms with a_i and rearrange to obtain

$$\begin{aligned} \mathcal{E} &\le \sum_i (\widetilde{a}_i - a_i)\frac{\widehat{\pi}_i - \pi^\star_{\alpha,i}}{\widehat{\pi}_i \pi^\star_{\alpha,i}} \\ &\le \eta_B \sum_i \frac{\sqrt{a_i}|\widehat{\pi}_i - \pi^\star_{\alpha,i}|}{\sqrt{\widehat{\pi}_i}\pi^\star_{\alpha,i}} \frac{\sqrt{a_i}}{\sqrt{\widehat{\pi}_i}} \\ &\le \eta_B \sqrt{D\, S(\widehat{\pi})} \le \eta_B \sqrt{\mathcal{E}\, S(\widehat{\pi})}. \end{aligned}$$

The third line is Cauchy–Schwarz applied to the two factors in the preceding sum. If $\mathcal{E} = 0$, the first upper bound in Eq. A.93 holds with zero left-hand side. Otherwise, square the inequality and divide by $\mathcal{E}$:

$$\mathcal{E} \le \eta_B^2 S(\widehat{\pi}) = \eta_B^2[S(\pi^\star_\alpha) + \mathcal{E}], \qquad \mathcal{E} \le \frac{\eta_B^2}{1-\eta_B^2} S(\pi^\star_\alpha).$$

Division is valid since $1 - \eta_B^2 > 0$. The surrogate difference is $\mathcal{E}/(2n^2)$. Finally, the vector with every entry B/n is feasible, giving

$$S(\pi^\star_\alpha) \le \frac{n}{B}\sum_i a_i \le \frac{n^2 a_{\max}}{B}.$$

Substitution proves the last bound. □

Coefficients near zero Relative accuracy is restrictive when exact scores are small; in particular, a positive coefficient floor cannot satisfy it at $a_i = 0$. For nonnegative a_i, an additive bound remains available. The approximate-score objective comparison in the preceding proof gives

$$\begin{aligned} \widetilde{\mathcal{R}}(\widehat{\pi}) - \widetilde{\mathcal{R}}(\pi^\star_\alpha) &\le \frac{1}{2n^2}\sum_i |\widetilde{a}_i - a_i| \left|\frac{1}{\pi^\star_{\alpha,i}} - \frac{1}{\widehat{\pi}_i}\right| \\ &\le \frac{1}{2n^2\alpha}\sum_i |\widetilde{a}_i - a_i| = \frac{1}{2cB}\frac{1}{n}\sum_i |\widetilde{a}_i - a_i|, \end{aligned} \tag{A.95}$$

where the equality uses $\alpha = cB/n$. The second inequality follows because both reciprocal probabilities belong to $[1, 1/\alpha]$. This bound is linear in average absolute coefficient error.

Corollary A.3 (Near-optimal risk with optimization and score errors)**.** *Suppose the conditions of Proposition 3 hold separately for $\widehat{\pi}$ and π_{opt}, with the same bound $\rho_B = O(B^{-1/2})$ and constants independent of B. Assume $0 < a_i \le a_{\max} = O(1)$ and Eq. A.92. Then*

$$\begin{aligned} &0 \le \mathcal{R}(\widehat{\pi}) - \mathcal{R}(\pi_{\mathrm{opt}}) \\ &= O\left(\frac{\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2}{B} + B^{-3/2} + \big(\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2 + B^{-1/2}\big)\rho_B + \rho_B^2 + \frac{\eta_B^2}{B} \right). \end{aligned} \tag{A.96}$$

Proof. Add and subtract the surrogate at the two designs, including the same C_B in each corrected difference:

$$\begin{aligned} \mathcal{R}(\widehat{\pi}) - \mathcal{R}(\pi_{\mathrm{opt}}) \le\ & |\mathcal{R}(\widehat{\pi}) - \widetilde{\mathcal{R}}(\widehat{\pi}) - C_B| \\ & + \widetilde{\mathcal{R}}(\widehat{\pi}) - \widetilde{\mathcal{R}}(\pi_\alpha^\star) \\ & + |\mathcal{R}(\pi_{\mathrm{opt}}) - \widetilde{\mathcal{R}}(\pi_{\mathrm{opt}}) - C_B|. \end{aligned}$$

Here surrogate optimality replaces its value at π_{opt} by its no larger value at $\pi_\alpha^\star$. The first and third terms each satisfy Eq. A.33. Proposition 12 bounds the middle term by

$$\frac{a_{\max}}{2B} \frac{\eta_B^2}{1-\eta_B^2} = O(1) B^{-1} O(\eta_B^2) = O(\eta_B^2/B),$$

since $1 - \eta_B^2 \ge 1 - \eta_0^2 > 0$. Adding the three bounds proves Eq. A.96; feasibility and optimality for the common training procedure give nonnegative regret. □

Rates and exact limits With $\rho_B = O(B^{-1})$, the residual terms contribute

$$\|\nabla_\theta F_{\mathrm{val}}(\theta_0)\|_2 O(B^{-1}), \qquad O(B^{-1/2})O(B^{-1}) = O(B^{-3/2}), \qquad O(B^{-1})^2 = O(B^{-2}).$$

The resulting rate is the main theorem's rate plus $O(\eta_B^2/B)$. If the validation gradient norm is $O(B^{-1/2})$ and $\eta_B = O(B^{-1/4})$, the remaining terms have orders $O(B^{-1/2})B^{-1} = O(B^{-3/2})$, $O(B^{-3/2})$, and $O(B^{-1/4})^2 B^{-1} = O(B^{-3/2})$. Their sum is $O(B^{-3/2})$. When $\eta_B = 0$, $\widehat{\pi} = \pi_\alpha^\star$ and the score term vanishes. When both optimization residuals are zero, $C_B = 0$ and $\rho_B = 0$ yields the exact-stationarity approximate-score result. These statements are conditional on the residual and stability bounds, rather than convergence guarantees for a fixed number of fine-tuning steps.

An alternative to a small validation gradient Corollary A.3 bounds each approximation error separately in absolute value, which is why it needs a small validation gradient. A signed comparison avoids this. Assume the conditions of Corollary A.1 hold separately for $\widehat{\pi}$ and π_{opt} under the same retraining rule, with constants uniform along the problem sequence. Let $\widehat{\pi}$ be the approximate-score design above, with both exact and effective coefficients evaluated at $\hat{\theta}$. Assume $a_i > 0$ for all i, $a_{\max} := \max_i a_i = O(1)$, and Eq. A.92, with $0 \le \eta_B \le \eta_0 < 1$ for a fixed η_0. For either $\pi \in \{\widehat{\pi}, \pi_{\mathrm{opt}}\}$, the exact decomposition in Eq. A.28 and the moments proved in Corollary A.1 give

$$\mathcal{R}(\pi) - \widetilde{\mathcal{R}}(\pi) = \nabla_\theta F_{\mathrm{val}}(\hat{\theta})^\top \mathbb{E}_W e_\pi + O(B^{-3/2}). \tag{A.97}$$

Indeed, the cross term, squared-error term, and Taylor remainder have orders $O(B^{-1/2})O(B^{-1}) = O(B^{-3/2})$, $O(B^{-2})$, and $O(B^{-3/2})$; their sum is $O(B^{-3/2})$. Proposition 12 gives

$$\begin{aligned} 0 &\le \widetilde{\mathcal{R}}(\widehat{\pi}) - \widetilde{\mathcal{R}}(\pi_\alpha^\star) \\ &\le \frac{a_{\max}}{2B} \frac{\eta_B^2}{1-\eta_B^2} = O(\eta_B^2/B). \end{aligned}$$

Here $1 - \eta_B^2 \ge 1 - \eta_0^2 > 0$, so the bound has order $O(1)B^{-1}O(\eta_B^2) = O(\eta_B^2/B)$. Subtracting Eq. A.97 at the two designs and using $\widetilde{\mathcal{R}}(\pi_\alpha^\star) \le \widetilde{\mathcal{R}}(\pi_{\mathrm{opt}})$ shows that the additional condition

$$\left| \nabla_\theta F_{\mathrm{val}}(\hat{\theta})^\top (\mathbb{E}_W e_{\widehat{\pi}} - \mathbb{E}_W e_{\pi_{\mathrm{opt}}}) \right| = O(B^{-3/2})$$

also gives regret $O(B^{-3/2} + \eta_B^2/B)$, without a small validation gradient. Each expectation here uses its own design. HT centers $\Delta_\pi^{\mathrm{lin}}$ but does not establish this cancellation of nonlinear retraining biases.

A quadratic check of the cancellation Consider a diagnostic quadratic training family with a fixed, unsampled curvature $H \succ 0$ and sampled linear terms:

$$F_{\mathrm{tr},\pi}(\theta_0 + u; W) = \tfrac{1}{2} u^\top H u + (r_0 + g_\pi^{\mathrm{c}})^\top u + \text{a term independent of } u.$$

This is a separate example with deterministic curvature, rather than a claim that HT sampling preserves per-example Hessians. With g_π^{c} induced by the sampled linear terms, $\overline{H}_\pi = H$, and the output identity is $\Delta_\pi = d_0 + \Delta_\pi^{\mathrm{lin}} + H^{-1} r_\pi$. For a quadratic validation loss the Taylor remainder is zero. With exact retraining, $r_\pi = 0$, so Eq. A.38 gives the exact equality $\mathcal{R}(\pi) = \widetilde{\mathcal{R}}(\pi) + C_B$ for every design. The common shift therefore cancels exactly in comparisons, even for a nonstationary reference point. With inexact retraining, $e_\pi - d_0 = H^{-1} r_\pi$ retains the cross term with $\Delta_\pi^{\mathrm{lin}}$; it cannot be discarded merely because the sampling perturbation has mean zero.

B Experimental Details and Additional Results

This section provides implementation details and additional results for the experiments in Section 4.

B.1 Implementation Details

All models in the one-epoch experiments are fine-tuned on NVIDIA 8×A100 GPUs.

We first LoRA fine-tune the base model on the full training set to obtain a reference model (Reference), whose parameters serve as $\hat{\theta}$ in Section 2. The sketched gradients, pseudo-label draws, and curvature estimates in Section 3.4 are all computed at $\hat{\theta}$, with CountSketch dimension $k = 4096$, and the validation objective of CAIS combines seven tasks. Given the resulting inclusion probabilities, we draw a Poisson sample with expected size B and LoRA fine-tune a new adapter from the base model on the selected subset. The retrained model is then evaluated and compared with the baselines under the same selection budget.

For all experiments, we use LoRA with rank 8, scaling factor $\alpha = 16$, and no dropout, applied to the query, key, value, and output projections (q, k, v, o) of the language model. The vision tower and the multimodal projector are frozen, while the language model is adapted through LoRA. Each retraining run initializes a new adapter from the base model rather than continuing from an existing one. We use AdamW with a learning rate of 10^{-4} and zero weight decay. The learning rate remains constant throughout training without warm-up. The per-device batch size is 1 with one gradient accumulation step on 8 GPUs, resulting in an effective global batch size of 8.

The random 20% / 80% split of each benchmark into validation and test examples uses seed 42. The three repeated runs in Section 4.2 use random seeds 101, 42, and 46, respectively.

B.2 Additional Results

Results under the 200K budget Table 6 reports results under a 200K-sample budget, complementing the 40K and 100K results in Section 4.2.

Table 6: Downstream evaluation results under a 200K-sample budget. Metric definitions follow Table 1; Rel. uses the Full row in this table as its 100% reference.

Method	MMBench-En	MMBench-Cn	MME-P	MME-C	AI2D	POPE-A	POPE-P	SQA-IMG	OCRBench	Rel.
Full	61.81	54.15	1319.64	268.75	54.01	83.40	93.72	69.39	14.70	100.00%
Uniform	60.11	54.04	1314.43	308.04	51.94	83.58	87.08	68.22	14.60	99.80%
Influence-Distillation	62.87	56.49	1345.25	273.21	50.85	82.56	94.16	66.73	14.20	99.56%
LESS	59.57	54.26	1125.28	297.77	49.47	83.43	80.22	66.54	14.80	96.27%
DiReCT	52.34	46.06	1408.97	265.63	41.21	83.36	92.94	55.14	14.70	92.25%
CAIS (Ours)	62.98	55.21	1386.76	262.95	52.71	84.43	91.49	66.79	14.60	99.87%

Measured selection and training cost Table 7 reports measured wall time for CAIS selection and one training epoch under the 40K selection budget, excluding Reference training and benchmark evaluation. This timing run uses a fresh base-model adapter, so its costs are reported separately from the multi-epoch continuation experiment in Section 4.3.

Candidate feature extraction dominates the measured runtime, accounting for 8.37 hours out of 9.03 hours, or approximately 92.8% of the total. Once the required features and curvature estimates are available, the selection stage takes only 37.70 seconds. This breakdown identifies candidate feature extraction as the main computational bottleneck and shows that solving the sampling design adds relatively little overhead.

Table 7: Measured runtime for CAIS selection and one training epoch under the 40K selection budget.

Stage	Time (s)	Time (h)
Candidate feature extraction	30142.93	8.3730
Validation curvature	482.78	0.1341
Selection	37.70	0.0105
Selected-data materialization	36.50	0.0101
Training configuration	2.20	0.0006
One-epoch training stage	1795.19	0.4987
Total (including setup and verification)	32498.60	9.0274

Distribution of the selected training set We compare the source composition of the full candidate pool with that of the 40K examples selected by CAIS. The candidate pool contains $624{,}610$ training records from five source datasets. The selected set contains $39{,}984$ records; the small deviation from $40{,}000$ is due to the realized Poisson draw, consistent with Section 2.2. Table 8 reports both the population composition and the composition of the selected set. Percentages in the two share columns are normalized by the corresponding set size, while the last column is the empirical selection rate within each source.

Table 8: Source distribution of the full candidate pool and the realized 40K CAIS sample. The selection rate is the fraction of each source retained by the sampler.

Source	Full count	Full share	Selected count	Selected share	Selection rate
COCO	364,100	58.29%	26,573	66.46%	7.30%
GQA	72,140	11.55%	3,917	9.80%	5.43%
OCR-VQA	80,000	12.81%	5,137	12.85%	6.42%
TextVQA	21,953	3.51%	2,232	5.58%	10.17%
Visual Genome	86,417	13.84%	2,125	5.31%	2.46%
Total	624,610	100.00%	39,984	100.00%	6.40%

The selected set is not a uniform source-level subsample. COCO accounts for 58.29% of the full pool but 66.46% of the selected set, corresponding to a $1.14\times$ enrichment. TextVQA is selected most aggressively: its share rises from 3.51% to 5.58%, and its within-source selection rate is 10.17%, approximately $1.59\times$ the overall rate. OCR-VQA is almost proportionally represented (12.81% in the pool versus 12.85% after selection), although its selection rate is still substantial at 6.42%. In contrast, Visual Genome is strongly downweighted: its share decreases from 13.84% to 5.31%, with only 2.46% of its records retained. GQA is also moderately downweighted, from 11.55% to 9.80%.

These shifts indicate that the method allocates probability at the record level rather than enforcing a source-balanced quota. The over-representation of COCO and TextVQA, together with the reduced contribution of Visual Genome, is therefore an empirical consequence of the target-aware selection scores. The table should be read as a description of the realized sample composition; it does not by itself establish that a particular source causes a downstream benchmark improvement.