From Average Sensitivity to Small-Loss Regret Bounds under Random-Order Model

Shinsaku Sakaue
CyberAgent, National Institute of Informatics, RIKEN AIP
shinsaku.sakaue@gmail.com

Yuichi Yoshida
National Institute of Informatics
yyoshida@nii.ac.jp

Abstract

We study online learning in the random-order model, where the multiset of loss functions is chosen adversarially but revealed in a uniformly random order. By extending the *batch-to-online* transformation of Dong and Yoshida (2023), we show that if an offline algorithm enjoys a $(1+\varepsilon)$-approximation guarantee, an *average sensitivity* bound controlled by a function $\varphi(\varepsilon)$, and stability with respect to ε, then we can obtain a *small-loss* regret bound typically of order $\tilde{O}(\varphi^\star(\mathrm{OPT}_T))$, where $\varphi^\star$ is the concave conjugate of φ, OPT_T is the offline optimum over T rounds, and $\tilde{O}$ hides polylogarithmic factors in T. Our result refines their original $(1+\varepsilon)$-approximate regret guarantee and applies to a broad class of problems, including online k-means clustering and online low-rank approximation. We further apply our approach to online submodular function minimization using $(1\pm\varepsilon)$-cut sparsifiers of submodular hypergraphs, obtaining a small-loss regret bound of $\tilde{O}(n^3+n^{3/4}\mathrm{OPT}_T^{3/4})$, where n is the ground-set size; we also demonstrate its applicability to online ℓ_1 regression. Our work sheds light on the power of sparsification and related algorithmic techniques in achieving small-loss regret bounds in the random-order model, without requiring structural assumptions on loss functions, such as linearity or smoothness.

1 Introduction

The online learning literature has been shaped by two canonical regimes: the stochastic and adversarial regimes. However, the i.i.d. assumption of the stochastic regime is often restrictive, while guarantees in the adversarial regime can be overly conservative. The *random-order model* provides a compelling bridge between these extremes by allowing an adversary to pick the multiset of inputs while requiring that their order be drawn uniformly at random. This recovers the stochastic regime when the inputs themselves are generated i.i.d., while still allowing us to leverage random ordering to sidestep the pessimism of fully adversarial sequences, yielding better guarantees in many settings. The random-order model has a long history in competitive analysis of online algorithms [8, 20, 24, 33] and has recently received increasing attention in regret analysis [9, 22, 50].

Under favorable stochastic assumptions, Follow-the-Leader achieves logarithmic regret, whereas in the adversarial setting it may incur linear regret (see, e.g., Cesa-Bianchi and Lugosi [13, Sections 3.2 and 4.3]). Competing under adversarially selected inputs requires stability, as reflected in the design of Follow-the-Regularized-Leader [4, 49] and in the connections to differential privacy [2, 6, 23]. For the random-order model, Dong and Yoshida [19] identified the *average sensitivity*—the change in an algorithm's output under uniformly random deletions of its input—as the key stability measure. This quantity captures the robustness of algorithms to random perturbations and has been gaining attention across diverse contexts, including graph algorithms [55, 56], dynamic programming [35], and clustering [63]. Dong and Yoshida [19] showed that if a $(1+\varepsilon)$-approximate offline algorithm with a bounded average sensitivity is available, then simply feeding its output obtained on the data observed so far into the next prediction yields a $(1+\varepsilon)$-approximate regret bound controlled by the average sensitivity; they termed this procedure the *batch-to-online* transformation. Informally, their $(1+\varepsilon)$-approximate

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

regret bound takes the form "Learner's cumulative loss $-(1+\varepsilon)\mathrm{OPT}_T = O(\varphi(\varepsilon)\log T)$," where OPT_T is the offline optimal cumulative loss over T rounds. The function φ represents how the average sensitivity of the offline algorithm depends on the approximation accuracy ε, and it grows as ε approaches zero, e.g., $\varphi(\varepsilon) \asymp 1/\varepsilon^3$.

While their result offers a bridge from average sensitivity to online learning under the random-order model, it only applies to the $(1+\varepsilon)$-approximate regret. The approximate regret has traditionally served as a reasonable performance measure when even the offline optimum cannot be attained efficiently, as in the case of online submodular function maximization [42, 53]. By contrast, in the framework of Dong and Yoshida [19], approximation is used to ensure stability even when exact offline optimization is tractable, and the approximate-regret guarantee gives the learner an unreasonable advantage of $\varepsilon\mathrm{OPT}_T$. This observation raises a natural question: What form should a *non-approximate* (standard) regret bound take in the random-order model, and how does the average sensitivity of the offline algorithm shape that bound?

1.1 Our contribution

We extend the fixed-ε, $(1+\varepsilon)$-approximate-regret framework of Dong and Yoshida [19] to obtain a regret bound typically governed by $\tilde{O}(\varphi^\star(\mathrm{OPT}_T))$, where $\varphi^\star$ is the *concave conjugate* of φ and $\tilde{O}$ hides polylogarithmic factors in T (and problem-dependent parameters). For example, when $\varphi(\varepsilon) \asymp 1/\varepsilon^q$ for some $q > 0$, this yields a bound of $\tilde{O}(\mathrm{OPT}_T^{q/(q+1)})$. Regret bounds that depend on OPT_T (up to polylogarithmic factors) are called *small-loss* regret bounds (also known as first-order or $L^\star$ bounds), which are highly desirable as they automatically exploit benign input sequences with small OPT_T.

Our core idea is to set the approximation parameter ε_t separately for each round t to optimize the accuracy–stability trade-off exposed by our regret decomposition (see Section 3). Two quantities to be balanced are the approximation term $\varepsilon_t\mathrm{OPT}_t/t$ and the average-sensitivity term $\varphi(\varepsilon_t)/t$. In Section 4, we show that per-round optimization of ε_t is sufficient for obtaining a regret bound of $\tilde{O}(\varphi^\star(\mathrm{OPT}_T))$, even though OPT_T is unknown in advance.[1] This adaptive choice of ε_t introduces one additional requirement on the offline algorithm: in addition to the fixed-ε average-sensitivity bound used by Dong and Yoshida [19], the output must be stable when the approximation parameter changes according to our rule for choosing ε_t (Assumption 9). We verify this requirement for the offline algorithm families used in our applications.

Importantly, our framework does not rely on linearity or smoothness assumptions on the loss functions, unlike most prior small-loss regret bounds (see Section 1.2). It therefore yields small-loss regret bounds for a broad class of problems, including specific examples studied by Dong and Yoshida [19], such as online k-means clustering and online low-rank matrix approximation. Moreover, in Section 5, we demonstrate its applicability to online submodular function minimization [25, 27], which was not covered by Dong and Yoshida [19]. We obtain an $\tilde{O}\big(n^3 + n^{3/4}\mathrm{OPT}_T^{3/4}\big)$ regret bound, where n is the ground-set size. This showcases the advantage of our framework, as the continuous extension of submodular functions, known as the *Lovász extension*, is not smooth (nor self-bounding). At a technical level, we leverage $(1\pm\varepsilon)$-cut sparsifiers for *submodular hypergraphs* [34] to design an offline algorithm with low average sensitivity. As another example with non-smooth losses, Section 6 shows an instantiation for online ℓ_1 regression based on *Lewis-weight sampling* [46]. More broadly, Section 7 outlines a general recipe for deriving small-loss regret bounds from offline approximation algorithms based on widely studied algorithmic techniques, such as coresets, sketching, sampling, or sparsifiers, provided that these algorithms have bounded average sensitivity and parameter-change stability.

1.2 Related work

Small-loss regret bounds. In prediction with N experts, an $O(\sqrt{\mathrm{OPT}_T \log N} + \log N)$ regret bound is achievable [13, Section 2.4], where OPT_T here is the cumulative loss of the best expert. Small-loss regret bounds for limited-feedback settings (e.g., bandit or partial information) with explicitly enumerated action sets or linear loss functions have been widely studied [3, 5, 21, 28, 36, 37, 39, 41, 44, 58]. In online convex optimization, small-loss regret bounds have been obtained under smoothness—more precisely, under the *self-bounding* property [45, Definition 4.26]—of

[1] A standard way to adapt to such an unknown scale is a doubling trick. In the random-order model, however, extending the fixed-ε framework of Dong and Yoshida [19] via doubling is not straightforward. We discuss this obstruction in Appendix G.

loss functions [52, 62, 64]. For online portfolio selection, Tsai et al. [54] obtain a small-loss regret bound without smoothness, using a local-norm analogue of the self-bounding property. By contrast, we impose minimal assumptions on the decision space and loss functions, and focus on a general connection between average sensitivity of offline approximation algorithms and small-loss regret bounds under the random-order model. Among prior works, the closest to ours is Abernethy et al. [2], who derived small-loss regret bounds in the adversarial environment from a stability notion closely connected to differential privacy. Related subsequent works [11, 57] develop oracle-efficient approaches that obtain small-loss regret bounds by leveraging privacy-style stability notions. Our approach can be viewed as a random-order analogue that replaces such privacy-style stability requirements with average sensitivity. In particular, differential privacy is a worst-case stability notion, whereas average sensitivity is an average-case requirement and hence generally less restrictive [56, Section 1.5]. Another relevant work is Lykouris et al. [39], who obtained small-loss regret bounds in graph bandits via an approximate-regret analysis using a doubling trick. This shares a high-level analogy with our idea of extending the approximate-regret analysis of Dong and Yoshida [19].

Random-order model. In competitive analysis, the random-order model has a long history, from the classical secretary problem to subsequent studies across various settings [8, 20, 33]. We refer the reader to Gupta and Singla [24] for an overview. Recently, there has been growing interest in the random-order model in regret analysis. Garber et al. [22] established polylogarithmic regret bounds for smooth losses with bounded gradients under cumulative strong convexity; later, Sherman et al. [50] achieved optimal logarithmic regret bounds in this setting. Bernasconi et al. [9] developed a simulation-based approach for adapting stochastic algorithms to the random-order model; their guarantee for full-information prediction includes an additive $\tilde{O}(\sqrt{T})$ regret term.

2 Preliminaries

Let $[n] = \{1, \dots, n\}$ for $n \in \mathbb{Z}_{>0}$ and $[0] = \emptyset$. We consider an online learning problem over $T \in \mathbb{Z}_{>0}$ rounds. We work throughout with standard Borel spaces and measurable randomized algorithms. Let $\mathcal{X}$ be the input space, Θ a nonempty decision space, and $\ell\colon \Theta \times \mathcal{X} \to [0,1]$ a measurable loss function. At each round $t \in [T]$, the learner selects $\theta_t \in \Theta$, observes $x_t \in \mathcal{X}$, and incurs $\ell(\theta_t, x_t)$. We assume the random-order model: the multiset of T data points $\{x_1, \dots, x_T\} \in \mathcal{X}^T$ is chosen adversarially in advance,[2] but it is ordered uniformly at random. Let $\ell(\theta, X) = \sum_{x \in X} \ell(\theta, x)$ for any multiset X of data points in $\mathcal{X}$ and any $\theta \in \Theta$. For each $t \in [T]$, define the offline optimum as $\mathrm{OPT}_t = \min_{\theta \in \Theta} \sum_{i=1}^{t} \ell(\theta, x_i)$; we assume that, for every finite multiset X, the minimum $\min_{\theta \in \Theta} \ell(\theta, X)$ is attained and its value is measurable in X. Note that OPT_t for $t < T$ are random variables depending on the random order of $\{x_t\}_{t=1}^{T}$, while OPT_T is deterministic once the multiset $\{x_t\}_{t=1}^{T}$ is fixed. The following lemma, which also appears in Dong and Yoshida [19], follows from the random-order assumption; see Appendix A for a proof.

Lemma 1 (Bound on sum of average optimal values)**.** *It holds that* $\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{OPT}_t / t\right] \le \mathrm{OPT}_T$.

Let Alg be the learner's algorithm for computing $\theta_1, \dots, \theta_T \in \Theta$, which may be randomized. Denoting by $\mathbb{E}_{\mathsf{Alg}, \{x_t\}}$ the expectation over the randomness of Alg and the random order of $\{x_t\}_{t=1}^{T}$, the learner's performance is measured by the regret defined as $\mathcal{R}_T := \mathbb{E}_{\mathsf{Alg}, \{x_t\}}[\sum_{t=1}^{T} \ell(\theta_t, x_t)] - \mathrm{OPT}_T$.

In this paper, we assume an exact-real oracle model with unit-cost real arithmetic and comparisons for running-time guarantees; further oracle assumptions are stated where they are used.

Average sensitivity. To define average sensitivity, we first introduce the total variation distance.

Definition 2 (Total variation)**.** For probability measures P, Q defined on a measurable space $(\Omega, \mathcal{F})$, we define the total variation distance between P and Q as $\mathrm{TV}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$.

The average sensitivity of a randomized algorithm is measured by the mean total-variation distance of output distributions on two datasets that differ by a single point deleted uniformly at random. Formally, we use the following definition adopted in Yoshida and Ito [63] and Dong and Yoshida [19].

Definition 3 (Average sensitivity)**.** Let $t \in \mathbb{Z}_{>0}$ and $X \in \mathcal{X}^t$. Consider a randomized algorithm $\mathcal{A}$ defined on all finite labeled multisets over $\mathcal{X}$, including the empty multiset, and returning a random

[2] We use the abuse of treating multiset occurrences as labeled whenever a multiset is indexed, sampled from, or modified.

Algorithm 1: Batch-to-online transformation with time-varying approximation accuracy

Input: Offline algorithm family $\{\mathcal{A}_\varepsilon\}_{\varepsilon\in(0,1]}$ and parameter $\lambda > 0$;
Initialize $\theta_1 \in \Theta$ arbitrarily;
for $t = 1, \dots, T$ **do**
  Play θ_t, receive x_t, and incur $\ell(\theta_t, x_t)$;
  Compute $\varepsilon_t \in (0, 1]$ from $X_t = \{x_1, \dots, x_t\}$ observed so far and λ by the rule in (1);
  Obtain $\theta_{t+1} \in \Theta$ by running $\mathcal{A}_{\varepsilon_t}$ on X_t;

element of Θ. We assume that its output distribution depends measurably on the input and use $\mathcal{A}(X)$ to denote both the random output and its distribution. The average sensitivity of $\mathcal{A}$ on X is defined as

$$\frac{1}{t}\sum_{x\in X} \mathrm{TV}(\mathcal{A}(X), \mathcal{A}(X \setminus x)),$$

where $X \setminus x$ denotes the multiset obtained by removing one occurrence of x from X. For a function $\beta\colon \mathbb{Z}_{>0} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$, we say that the average sensitivity of $\mathcal{A}$ is upper bounded by β if the above quantity is at most $\beta(t)$ for every $t \in \mathbb{Z}_{>0}$ and $X \in \mathcal{X}^t$.

Approximation accuracy and average sensitivity. Keeping an algorithm's average sensitivity low typically comes at the expense of approximation accuracy. In this study, we focus on algorithms that exhibit the following trade-off between approximation accuracy and average sensitivity.

Assumption 4 (Offline average-stable approximation)**.** Let $t \in [T]$ and $X \in \mathcal{X}^t$. For any fixed $\varepsilon \in (0, 1]$, the offline randomized algorithm, denoted by $\mathcal{A}_\varepsilon$, returns a $(1+\varepsilon)$-approximate solution in expectation, up to an additive $O(t/T)$ term (where the hidden constant may depend on problem-dependent parameters but not on t, T, ε, or X), and its average sensitivity is at most $\varphi(\varepsilon)/(2t)$, i.e.,

$$\mathbb{E}_{\mathcal{A}_\varepsilon}[\ell(\mathcal{A}_\varepsilon(X), X)] \leq (1+\varepsilon)\min_{\theta\in\Theta} \ell(\theta, X) + O\left(\frac{t}{T}\right) \text{ and } \frac{1}{t}\sum_{x\in X} \mathrm{TV}(\mathcal{A}_\varepsilon(X), \mathcal{A}_\varepsilon(X \setminus x)) \leq \frac{\varphi(\varepsilon)}{2t},$$

where $\varphi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ is a proper lower-semicontinuous function with effective domain $\mathrm{dom}(\varphi) \coloneqq \{\, \varepsilon \in \mathbb{R}_{\geq 0} \,:\, \varphi(\varepsilon) < +\infty \,\}$ contained in $(0, 1]$, and such that $\lim_{\varepsilon\downarrow 0} \varphi(\varepsilon) = +\infty$.

Typically, $\varphi(\varepsilon)$ is a decreasing function on $(0, 1]$, e.g., $\varphi(\varepsilon) \asymp 1/\varepsilon^3$, and thus a better approximation (smaller ε) increases the average sensitivity bound $\varphi(\varepsilon)/(2t)$, reflecting the trade-off between approximation accuracy and average sensitivity. The $O(t/T)$ term in the approximation guarantee is non-essential and is included only for consistency with high-probability guarantees of offline algorithms. The role of Assumption 4 is to isolate how ε and t affect the average sensitivity bound, which is indeed satisfied in all examples discussed later.[3] In Section 4.2, we also introduce an additional condition (Assumption 9) on offline algorithms to handle time-varying approximation accuracy.

3 Batch-to-online transformation with time-varying approximation accuracy

We introduce the batch-to-online transformation. As described in Algorithm 1, at each round t, the offline algorithm $\mathcal{A}_{\varepsilon_t}$ is applied to the dataset X_t observed so far to obtain θ_{t+1}. A crucial difference from Dong and Yoshida [19] is that we allow the approximation parameter ε_t to vary over rounds. This time-varying view yields a new regret decomposition, exposing the per-round accuracy–stability trade-off and the cost of changing ε_t over time.

Proposition 5 (Regret decomposition)**.** *Suppose that the parameters ε_t used by Algorithm 1 are specified by a measurable rule $Y \mapsto \varepsilon_Y$ for any finite multiset Y, so that $\varepsilon_t = \varepsilon_{X_t}$ holds for every t; let $\varepsilon_\emptyset \in (0, 1]$ be an arbitrary fixed value. For $\theta_1, \dots, \theta_T \in \Theta$ obtained by Algorithm 1, the regret $\mathcal{R}_T$ is upper bounded by the following two sums up to an additive constant:*[4]

$$\mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\left(\frac{\varepsilon_t}{t}\mathrm{OPT}_t + \frac{\varphi(\varepsilon_t)}{t}\right)}_{\text{accuracy–stability trade-off}}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\frac{2}{t}\sum_{x\in X_t} \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t \setminus x)\Big)}_{\text{parameter-change cost}}\right] + O(1).$$

[3] Even when an average sensitivity bound $\beta(t)$ is given as $\tilde{\varphi}(t, \varepsilon)/(2t)$ for some $\tilde{\varphi}$ increasing in t, setting $\varphi(\varepsilon) = \tilde{\varphi}(T, \varepsilon)$ yields a valid upper bound. Typically, such $\tilde{\varphi}(t, \varepsilon)$ is logarithmic in t, and this change affects the bound only logarithmically.

[4] Terms involving the empty input contribute only an additive $O(1)$.

We defer the proof to Appendix B, but the intuition is as follows. The output θ_{t+1} returned by the $(1+\varepsilon_t)$-approximation algorithm satisfies $\ell(\theta_{t+1}, X_t)/t \lesssim (1+\varepsilon_t)\mathrm{OPT}_t/t$, and evaluating it on the fresh loss $\ell(\cdot, x_{t+1})$ introduces an additive error governed by the average sensitivity plus the cost of using $\varepsilon_t = \varepsilon_{X_t}$ on $X_t \setminus x$ instead of $\varepsilon_{X_t \setminus x}$, hence $\mathbb{E}[\ell(\theta_{t+1}, x_{t+1})] \lesssim \mathbb{E}[(1+\varepsilon_t)\mathrm{OPT}_t/t + \varphi(\varepsilon_t)/t] +$ parameter-change cost. Summing over t and applying Lemma 1 yield the above regret decomposition.

Observation from the time-varying view. Allowing the approximation parameter ε_t to vary over rounds reveals a more informative structure than the fixed-ε presentation of Dong and Yoshida [19]: the accuracy–stability trade-off term takes the form of a *per-round* trade-off between the two terms, $\frac{\mathrm{OPT}_t}{t}\varepsilon_t$ and $\frac{\varphi(\varepsilon_t)}{t}$. Although the target quantity OPT_T is not known in advance, the realized value of OPT_t is determined by X_t. In addition, as we will see later, this trade-off term typically dominates the parameter-change cost. Therefore, a natural strategy is: choose ε_t by optimizing the per-round trade-off, which is the perspective underlying the next section.

4 Small-loss regret bound via time-varying approximation accuracy

We present how to obtain a small-loss regret bound by bounding the two sums in the regret decomposition in Proposition 5. The following concave conjugate of φ plays a central role in our analysis.

Definition 6 (Concave conjugate)**.** Let $\varphi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ be a proper lower-semicontinuous function with $\mathrm{dom}(\varphi) \subseteq (0, 1]$ and $\lim_{\varepsilon \downarrow 0} \varphi(\varepsilon) = +\infty$. Its concave conjugate $\varphi^\star\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is $\varphi^\star(u) \coloneqq \inf_{\varepsilon \geq 0}\{u\varepsilon + \varphi(\varepsilon)\}$ for $u \geq 0$. For $u > 0$, this infimum is attained; let $\varepsilon_{\min}(u)$ denote a minimizer selected by a fixed tie-breaking rule.[5]

For example, if $\varphi(\varepsilon) \asymp \varepsilon^{-q}$ for some fixed $q > 0$, then $\varphi^\star(u) \asymp 1 + u^{q/(q+1)}$ holds. We will use the following basic properties of $\varphi^\star$ and $\varepsilon_{\min}$; see Appendix C for the proof.

Lemma 7. *The function $\varphi^\star$ is finite-valued, non-decreasing, and concave on $\mathbb{R}_{\geq 0}$. The selected minimizer $\varepsilon_{\min}\colon \mathbb{R}_{>0} \to (0, 1]$ is non-increasing and, for every $u > 0$, is a supergradient of $\varphi^\star$ at u, i.e., $\varepsilon_{\min}(u)(u - v) \leq \varphi^\star(u) - \varphi^\star(v)$ for every $v \geq 0$.*

4.1 Choosing approximation accuracy to optimize per-round trade-off

We show how to choose ε_t. The core idea is to set ε_t to optimize the accuracy–stability trade-off, or $\mathrm{OPT}_t\varepsilon_t + \varphi(\varepsilon_t)$. We derive a useful bound on the optimized trade-off terms by leveraging the properties of $\varphi^\star$. Note that, while the following choice of ε_t involves OPT_t, which varies with the random order, the bound holds for every realization of the randomness. Here, the rule assumes that the exact optimum value OPT_t can be computed from the observed multiset $X_t = \{x_s\}_{s=1}^t$ when choosing ε_t.

Lemma 8 (Per-round trade-off optimization)**.** *Let $\lambda > 0$ and consider the following choice of ε_t:*

$$\varepsilon_t = \varepsilon_{\min}(\mathrm{OPT}_t + \lambda) \in \arg\min\{\,(\mathrm{OPT}_t + \lambda)\varepsilon + \varphi(\varepsilon) \;:\; \varepsilon \geq 0\,\} \qquad \text{for } t = 1, \ldots, T. \tag{1}$$

Writing $H_T = \sum_{s=1}^T \frac{1}{s}$, the accuracy–stability trade-off part in Proposition 5 is bounded as follows:

$$\sum_{t=1}^T \left(\frac{\mathrm{OPT}_t}{t}\varepsilon_t + \frac{\varphi(\varepsilon_t)}{t}\right) \leq H_T \varphi^\star\left(\frac{\sum_{t=1}^T \mathrm{OPT}_t/t}{H_T} + \lambda\right).$$

Proof. By the definition of $\varepsilon_t = \varepsilon_{\min}(\mathrm{OPT}_t + \lambda)$, we have $(\mathrm{OPT}_t + \lambda)\varepsilon_t + \varphi(\varepsilon_t) = \varphi^\star(\mathrm{OPT}_t + \lambda)$. Rearranging the terms and using $\lambda > 0$ and $\varepsilon_t \geq 0$ yield $\mathrm{OPT}_t\varepsilon_t + \varphi(\varepsilon_t) = \varphi^\star(\mathrm{OPT}_t + \lambda) - \lambda\varepsilon_t \leq \varphi^\star(\mathrm{OPT}_t + \lambda)$. Dividing by t and summing over $t = 1, \ldots, T$ give

$$\sum_{t=1}^T \left(\frac{\mathrm{OPT}_t}{t}\varepsilon_t + \frac{\varphi(\varepsilon_t)}{t}\right) \leq \sum_{t=1}^T \frac{\varphi^\star(\mathrm{OPT}_t + \lambda)}{t} = H_T \sum_{t=1}^T \frac{t^{-1}}{H_T}\varphi^\star(\mathrm{OPT}_t + \lambda).$$

[5] Since the divergence at zero excludes a sufficiently small interval $(0, \delta)$ from the relevant sublevel set of $u\varepsilon + \varphi(\varepsilon)$, the minimization reduces to the compact interval $[\delta, 1]$; therefore, lower-semicontinuity gives a minimizer. We also require $\varepsilon_{\min}$ to be Borel-measurable; in the applications, this is immediate as the minimizer is unique except possibly at the endpoint. For specific profiles of φ used below, computing $\varepsilon_{\min}$ is treated as a unit-cost operation in the time-complexity bounds.

Since $\varphi^\star$ is concave, Jensen's inequality applied to the probability weights $(t^{-1}/H_T)_{t=1}^T$ gives

$$H_T \sum_{t=1}^{T} \frac{t^{-1}}{H_T} \varphi^\star(\mathrm{OPT}_t + \lambda) \le H_T \varphi^\star\left(\sum_{t=1}^{T} \frac{t^{-1}}{H_T} (\mathrm{OPT}_t + \lambda) \right) = H_T \varphi^\star\left(\frac{\sum_{t=1}^{T} \mathrm{OPT}_t/t}{H_T} + \lambda \right).$$

Combining the above inequalities completes the proof. □

We introduce $\lambda > 0$ as a small shift to ensure, together with $\mathrm{dom}(\varphi) \subseteq (0, 1]$ and $\lim_{\varepsilon \downarrow 0} \varphi(\varepsilon) = +\infty$, that ε_t is well defined in $(0, 1]$ even when $\mathrm{OPT}_t = 0$.

4.2 Bounding parameter-change cost

To control the parameter-change cost part in Proposition 5, we need an additional condition tailored to the parameter selection rule (1). This is indeed satisfied in all applications considered in this paper.

Assumption 9 (Parameter-change stability)**.** Fix $\lambda > 0$ and consider the selection rule of $\varepsilon_{\min}$ in (1). For any multiset Y, we write $\mathrm{OPT}(Y) := \min_{\theta \in \Theta} \ell(\theta, Y)$ and $\varepsilon_Y := \varepsilon_{\min}(\mathrm{OPT}(Y) + \lambda)$, where $\mathrm{OPT}(\emptyset) = 0$ and $\varepsilon_\emptyset = \varepsilon_{\min}(\lambda)$. Under this setting, we assume that the offline algorithm family $\{\mathcal{A}_\varepsilon\}_{\varepsilon \in (0,1]}$ satisfies the following condition with constant $C_{\mathrm{pc}} \ge 0$ for any multiset X with $|X| \ge 2$:

$$\frac{1}{|X|} \sum_{x \in X} \mathrm{TV}\big(\mathcal{A}_{\varepsilon_X}(X \setminus x), \mathcal{A}_{\varepsilon_{X \setminus x}}(X \setminus x)\big) \le \frac{C_{\mathrm{pc}} \varepsilon_X}{2|X|} \sum_{x \in X} (\mathrm{OPT}(X) - \mathrm{OPT}(X \setminus x)).$$

Assumption 9 is for controlling the extra total variation caused by changing only the accuracy parameter from ε_X to $\varepsilon_{X \setminus x}$ on the same input $X \setminus x$.

Two-step sufficient condition. In the applications below, we can verify this condition through a common two-step argument. First, for a fixed input Y, the output distributions satisfy

$$\mathrm{TV}(\mathcal{A}_\varepsilon(Y), \mathcal{A}_\eta(Y)) = O\left(\varphi(\varepsilon) \frac{\eta - \varepsilon}{\varepsilon} \right)$$

for $0 < \varepsilon \le \eta \le 1$. Second, the rule $\varepsilon_Y = \varepsilon_{\min}(\mathrm{OPT}(Y) + \lambda)$, together with a regularity condition of φ in Appendix J, converts the relative parameter change into $\varepsilon_X(\mathrm{OPT}(X) - \mathrm{OPT}(X \setminus x))$, and averaging over $x \in X$ ensures the requirement in Assumption 9.

The following lemma is a direct consequence of Assumption 9; see Appendix D for the proof.

Lemma 10 (Controlling parameter-change cost)**.** *Let $\mathrm{OPT}_0 = 0$ and assume Assumption 9. Then, the parameter-change cost part in Proposition 5 is bounded as follows:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{2}{t} \sum_{x \in X_t} \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x) \Big) \right] \le C_{\mathrm{pc}} \mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t (\mathrm{OPT}_t - \mathrm{OPT}_{t-1}) \right] + O(1).$$

4.3 Small-loss regret bound

Lemmas 8 and 10 combined with Proposition 5 yield the following small-loss regret bound.

Theorem 11 (Small-loss regret bound)**.** *Under the random-order model and Assumption 4, fix $\lambda > 0$ and suppose that Assumption 9 holds for this λ. Then, Algorithm 1 with ε_t given by (1) achieves*

$$\mathcal{R}_T \le H_T \varphi^\star\left(\frac{\mathrm{OPT}_T}{H_T} + \lambda \right) + C_{\mathrm{pc}}(\varphi^\star(\mathrm{OPT}_T + \lambda) - \varphi^\star(\lambda)) + O(1).$$

Proof. First, we bound the accuracy–stability trade-off part in Proposition 5 via Lemma 8. Writing $u_T = (\sum_{t=1}^{T} \mathrm{OPT}_t/t)/H_T + \lambda$, Jensen's inequality for concave $\varphi^\star$ gives $\mathbb{E}[\varphi^\star(u_T)] \le \varphi^\star(\mathbb{E}[u_T])$. Since $\mathbb{E}[u_T] \le \mathrm{OPT}_T/H_T + \lambda$ follows from Lemma 1 and $\varphi^\star$ is non-decreasing, Lemma 8 implies that the accuracy–stability trade-off part is at most $H_T \varphi^\star(\mathrm{OPT}_T/H_T + \lambda)$. We then upper bound the parameter-change cost part via Lemma 10. Let $a_t = \mathrm{OPT}_t + \lambda$ for $t \in [T]$ and $a_0 = \lambda$. It holds that $\varepsilon_t(\mathrm{OPT}_t - \mathrm{OPT}_{t-1}) = \varepsilon_{\min}(a_t)(a_t - a_{t-1}) \le \varphi^\star(a_t) - \varphi^\star(a_{t-1})$ thanks to the supergradient property of $\varepsilon_{\min}$ in Lemma 7. Telescoping the sum over t, taking expectation, and using Lemma 10 yield a bound of $C_{\mathrm{pc}}(\varphi^\star(\mathrm{OPT}_T + \lambda) - \varphi^\star(\lambda))$ on the parameter-change cost part. Finally, summing the bounds on the two parts completes the proof. □

Below, we instantiate this theorem for a commonly used form of φ; see Appendix E for the derivation.

Corollary 12. *Let $q > 0$ be a constant. Assume $\varphi(\varepsilon) = C_1 \varepsilon^{-q} \log(\mathrm{e} + C_2/\varepsilon)$ for $\varepsilon \in (0, 1]$ and $\varphi(\varepsilon) = +\infty$ for $\varepsilon \notin (0, 1]$ for some $C_1 \geq 1$ and $C_2 \geq 0$. Then, for any $\lambda \in (0, H_T^{-(q+1)/q}]$ such that Assumption 9 holds, Theorem 11 yields $\mathcal{R}_T = \tilde{O}((1 + C_{\mathrm{pc}})(C_1 + C_1^{1/(q+1)} \mathrm{OPT}_T^{q/(q+1)}))$.*

In the above bound, the dependence on C_2 is logarithmic and thus hidden by $\tilde{O}$; see Appendix E for the explicit bound. Note that C_1, C_2, and C_{pc} may depend on problem-specific parameters and T.

Recovering the $(1 + \varepsilon)$-approximate regret bound. We derive the $(1 + \varepsilon)$-approximate regret guarantee of Dong and Yoshida [19] from Theorem 11. Under $C_{\mathrm{pc}} = O(1)$, fix $\varepsilon \in (0, 1]$ with $\varphi(\varepsilon/(1 + C_{\mathrm{pc}})) < +\infty$, and choose $\lambda \in (0, 1/(\varepsilon H_T)]$ such that Assumption 9 holds. Set $u := \mathrm{OPT}_T/H_T + \lambda$. Since $H_T \geq 1$, we have $\mathrm{OPT}_T + \lambda \leq H_T u$; therefore, nonnegativity, monotonicity, and concavity of $\varphi^\star$ with $\varphi^\star(0) \geq 0$ imply

$$0 \leq \varphi^\star(\mathrm{OPT}_T + \lambda) - \varphi^\star(\lambda) \leq \varphi^\star(\mathrm{OPT}_T + \lambda) \leq \varphi^\star(H_T u) \leq H_T \varphi^\star(u).$$

Thus, Theorem 11 gives $\mathcal{R}_T \leq (1 + C_{\mathrm{pc}}) H_T \varphi^\star(u) + O(1)$. Since $\varphi^\star(u) = \inf_{\eta \geq 0}\{u\eta + \varphi(\eta)\}$, choosing $\eta = \varepsilon/(1 + C_{\mathrm{pc}})$ gives

$$\varphi^\star(u) \leq \varepsilon u/(1 + C_{\mathrm{pc}}) + \varphi(\varepsilon/(1 + C_{\mathrm{pc}})) = \varepsilon(\mathrm{OPT}_T/H_T + \lambda)/(1 + C_{\mathrm{pc}}) + \varphi(\varepsilon/(1 + C_{\mathrm{pc}})).$$

Combining this with $\varepsilon H_T \lambda \leq 1$ yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\theta_t, x_t)\right] - (1 + \varepsilon)\mathrm{OPT}_T = \mathcal{R}_T - \varepsilon \mathrm{OPT}_T \leq (1 + C_{\mathrm{pc}}) H_T \varphi\left(\frac{\varepsilon}{1 + C_{\mathrm{pc}}}\right) + O(1).$$

This recovers the $(1 + \varepsilon)$-approximate regret bound of Dong and Yoshida [19] up to $O(1)$, with $\varphi(\varepsilon)$ replaced by $(1 + C_{\mathrm{pc}})\varphi(\varepsilon/(1 + C_{\mathrm{pc}}))$. Thus, our bound refines their fixed-ε result by tuning ε to optimize the regret bound without knowing OPT_T; note that a naive doubling does not apply in this random-order analysis (see Appendix G).

Inconsistency bound. In addition to regret, Dong and Yoshida [19] considered *inconsistency* Inc_T, the expected number of changes in the maintained outputs between consecutive rounds. This notion is relevant when changing a deployed decision incurs switching, communication, or recomputation costs. Using independent reruns of Algorithm 1 is not sufficient for controlling this quantity. However, as shown in Appendix F, if consecutive output distributions are implemented through probability transformation [63, Definition 4.1 and Lemma 4.5], we can prove $\mathrm{Inc}_T \lesssim H_T \varphi^\star(\mathrm{OPT}_T/H_T + \lambda) + C_{\mathrm{pc}}(\varphi^\star(\mathrm{OPT}_T + \lambda) - \varphi^\star(\lambda))$ (see Theorem F.2 for details), while, conditional on the ordered history $(x_1, \ldots, x_t)$, the post-update output θ_{t+1} has law $\mathcal{A}_{\varepsilon_t}(X_t)$ and hence the regret guarantee is unchanged. For the function φ in Corollary 12, this recovers the fixed-ε inconsistency bound of $O(\varphi(\varepsilon) \log T)$ obtained by Dong and Yoshida [19] (see Corollary F.3).

5 Application to online submodular function minimization

This section shows how our framework applies to online submodular function minimization [25]. Let V be a finite ground set of size $n \in \mathbb{Z}_{>0}$ and $\Theta = 2^V$ the decision set. We write $\ell_t(\cdot)$ for $\ell(\cdot, x_t)$ for convenience, and we assume that each $\ell_t \colon 2^V \to [0, 1]$ is a submodular function, i.e., $\ell_t(S_1) + \ell_t(S_2) \geq \ell_t(S_1 \cup S_2) + \ell_t(S_1 \cap S_2)$ for any $S_1, S_2 \subseteq V$. In the random-order model, the multiset of functions ℓ_t is chosen by an adversary but they arrive in a uniformly random order. We sequentially select $S_t \in \Theta$ to minimize the regret against the optimal value $\mathrm{OPT}_T = \min_{S \in \Theta} \sum_{t=1}^{T} \ell_t(S)$.

5.1 Offline algorithm

Our offline algorithm uses the importance-sampling method underlying the submodular-hypergraph sparsification result of Kenneth and Krauthgamer [34]; see Appendix K for the construction and its analysis.

Let $H = (U, E, \{g_e\}_{e \in E})$ be a submodular hypergraph, where U is a finite vertex set, E is a multiset of hyperedges on U, and each $e \in E$ has a non-negative submodular splitting function $g_e \colon 2^e \to \mathbb{R}_{\geq 0}$ with $g_e(\emptyset) = 0$. We call $H' = (U, E', \{g'_e\}_{e \in E'})$ a *reweighted subgraph* of H if $E' \subseteq E$ and for each

$e \in E'$, there exists $s_e > 0$ such that $g'_e \equiv s_e g_e$. We define the cut value as $\mathrm{cut}_H(A) = \sum_{e \in E} g_e(A \cap e)$ for $A \subseteq U$. Kenneth and Krauthgamer [34, Theorem 1.4] show that, for any $\varepsilon > 0$, H admits a reweighted subgraph H' with $|E'| = \tilde{O}(\varepsilon^{-2}|U|^3)$ satisfying

$$(1-\varepsilon) \cdot \mathrm{cut}_H(A) \le \mathrm{cut}_{H'}(A) \le (1+\varepsilon) \cdot \mathrm{cut}_H(A) \qquad \text{for all} \quad A \subseteq U. \tag{2}$$

For the randomized sampler used by our offline algorithm, Appendix K proves the two guarantees separately: for every $\delta \in (0,1)$, (2) holds with probability at least $1-\delta$, while its output satisfies $\mathbb{E}[|E'|] = O\big(\varepsilon^{-2}(|U| + \log(1/\delta))(|U|^2+1)\big)$.

To apply this guarantee to online submodular function minimization, where $\ell_t(\emptyset)$ need not be zero, we use a one-vertex anchored lift: let $r \notin V$ be a dummy vertex, set $U = V^+ := V \cup \{r\}$, and, for each loss $\ell_t \colon 2^V \to [0,1]$, create a hyperedge $e_t = U$ with splitting function $g_{e_t} \colon 2^U \to \mathbb{R}_{\ge 0}$ defined by $g_{e_t}(\emptyset) = 0$ and $g_{e_t}(A) = \ell_t(A \cap V)$ for $A \neq \emptyset$. This gives valid non-negative submodular splitting functions (see Appendix K.1). With $E_t = \{e_1, \dots, e_t\}$ and $H_t^+ = (U, E_t, \{g_e\}_{e \in E_t})$, each original action $S \subseteq V$ is represented by the anchored cut $S \cup \{r\}$, for which $\mathrm{cut}_{H_t^+}(S \cup \{r\}) = \sum_{s=1}^t \ell_s(S)$. The offline algorithm constructs a $(1 \pm \varepsilon')$-cut sparsifier $\tilde{H}_t^+$ of H_t^+ for appropriate $\varepsilon' = \Theta(\varepsilon)$ and returns a minimizer of $S \mapsto \mathrm{cut}_{\tilde{H}_t^+}(S \cup \{r\})$ over $S \subseteq V$, using a fixed deterministic tie-breaking rule.

The main remaining task is to upper bound the average sensitivity of this offline algorithm. For a submodular hypergraph $H = (U, E, \{g_e\}_{e \in E})$, let $H - e$ denote the hypergraph obtained by removing $e \in E$ and its splitting function g_e from H. Consider applying the offline algorithm to H and $H - e$. Thanks to the data processing inequality, the deterministic minimization step does not increase the total variation distance between the two output distributions. Therefore, it suffices to analyze the average sensitivity of the sparsifier construction step. In Appendix K, we prove the following bound.

Proposition 13. *Let $\varepsilon \in (0,1]$. For a submodular hypergraph $H = (U, E, \{g_e\}_{e \in E})$ with $|E| \ge 1$, let $\mathcal{L}_\varepsilon(H)$ be the law of the $(1 \pm \varepsilon)$-cut sparsifier constructed by the method of Kenneth and Krauthgamer [34] (with weight perturbation in Appendix K.3), which satisfies (2) with probability at least $1 - (2T)^{-1}$. The average sensitivity of $\mathcal{L}_\varepsilon$ is bounded as $\sum_{e \in E} \mathrm{TV}(\mathcal{L}_\varepsilon(H), \mathcal{L}_\varepsilon(H-e))/|E| = O\big(\varepsilon^{-3}(|U|^3 + |U|^2 \log T)/|E|\big)$.*

Since the lifted hypergraph has $|U| = n+1$ vertices, this implies $\varphi(\varepsilon) = C_{\mathrm{sub}} \varepsilon^{-3}(n^3 + n^2 \log T)$ for some constant $C_{\mathrm{sub}} \ge 1$. Proposition K.5 also shows that Assumption 9 is satisfied with $C_{\mathrm{pc}} = O(1)$.

5.2 Regret analysis

Given the above offline approximation algorithm equipped with the average sensitivity bound and the parameter-change stability, we can apply Corollary 12 to obtain the following result for online submodular function minimization under the random-order model.

Theorem 14. *For online submodular function minimization under the random-order model, Algorithm 1 with $\lambda = H_T^{-4/3}$ runs in polynomial time and achieves $\mathcal{R}_T = \tilde{O}\big(n^3 + n^{3/4}\mathrm{OPT}_T^{3/4}\big)$.*

Proof. Recall that the multiset of t submodular functions $\ell_1, \dots, \ell_t$ corresponds to a submodular hypergraph H_t^+ on $n+1$ vertices with $|E_t| = t$ hyperedges. Then, Proposition 13 guarantees that the $(1+\varepsilon)$-approximate offline algorithm satisfies an average sensitivity bound with $\varphi(\varepsilon) = C_{\mathrm{sub}} \varepsilon^{-3}(n^3 + n^2 \log T)$. Moreover, Proposition K.5 establishes Assumption 9 with $C_{\mathrm{pc}} = O(1)$. Thus, Corollary 12 with $q = 3$, $C_1 = C_{\mathrm{sub}}(n^3 + n^2 \log T)$, $C_2 = 0$, and $C_{\mathrm{pc}} = O(1)$ proves the regret bound. The time-complexity claim follows from the implementation detailed in Appendix K, which runs in polynomial time in the stated exact-real value-oracle model. □

The regret bound in Theorem 14 can be nontrivial in the regime $T \gtrsim n^3$ up to polylogarithmic factors; otherwise, the trivial bound $\mathcal{R}_T \le T$ is better.

For online submodular function minimization in the adversarial regime, an $O(\sqrt{n\mathrm{OPT}_T} + n)$ regret bound is achievable by treating 2^n subsets as experts and applying the multiplicative weights update (see, e.g., Cesa-Bianchi and Lugosi [13, Section 2.4]).[6] However, this is prohibitively expensive,

[6] An idealized continuous-domain exponential-weights argument over the Lovász extension also suggests a possible route to small-loss bounds. However, the efficient random-walk tracking result of Narayanan and Rakhlin [40] is proved for linear Lipschitz losses and does not by itself yield a polynomial-time small-loss guarantee for the non-smooth Lovász-extension losses.

and therefore researchers have explored polynomial-time algorithms. Hazan and Kale [25] obtained a regret bound of $O(\sqrt{nT})$ in the adversarial regime with the online subgradient descent. Ito [27] obtained a best-of-both-worlds bound—$O(n/\Delta)$ in the stochastic regime with the suboptimality gap $\Delta > 0$ and $O(\sqrt{nT})$ in the adversarial regime—and extended it to the stochastic setting with adversarial corruptions. Here, Δ denotes the expected-loss gap between the optimal subset and the best alternative subset in the stochastic model. Unlike these prior results, we focus on the random-order model and obtain a small-loss regret bound. Notably, when $\mathrm{OPT}_T = O(1)$, our result achieves a regret bound that is polylogarithmic in T for fixed n even under the random-order model, without any assumption on the suboptimality gap Δ.

6 Application to online ℓ_1 regression

We next instantiate the framework for online ℓ_1 regression. Fix $d \in \mathbb{Z}_{>0}$. At each round t, the learner incurs the absolute loss $\ell(\theta_t, x_t) = |\langle a_t, \theta_t \rangle - b_t|$, where $x_t = (a_t, b_t) \in \mathbb{R}^d \times \mathbb{R}$, and computes $\theta_{t+1} \in \Theta \subseteq \mathbb{R}^d$ from $\{x_s\}_{s=1}^t$. We assume that Θ is nonempty, compact, and convex. For every finite sample $\{(a_i, b_i)\}_{i=1}^m$ and nonnegative weights $\omega_1, \dots, \omega_m$, we assume that an exact minimizer over Θ of $\sum_{i=1}^m \omega_i |\langle a_i, \theta \rangle - b_i|$ can be computed in polynomial time using a fixed tie-breaking rule. We also assume $\ell(\theta, x_t) \in [0, 1]$ for every $\theta \in \Theta$ and every adversarially chosen x_t, possibly after a known problem-dependent scaling. The data points arrive in a uniformly random order.

Offline algorithm. We use a $(1 + \varepsilon)$-approximate offline ℓ_1-regression algorithm based on Lewis-weight sampling [17, 46]; see Appendix L. It samples rows according to approximate Lewis weights and returns an exact minimizer of the sampled weighted ℓ_1 objective, which uniformly approximates the original objective with high probability. We can bound the average sensitivity of this algorithm by $\varphi(\varepsilon)/(2t)$ with $\varphi(\varepsilon) = C_{\ell_1} d\varepsilon^{-3} \log(\mathrm{e} + dT/\varepsilon)$ for some constant $C_{\ell_1} \geq 1$ and verify Assumption 9 with $C_{\mathrm{pc}} = O(1)$; see Propositions L.7 and L.8, respectively.

Regret analysis. Given the offline algorithm with $\varphi(\varepsilon) = C_{\ell_1} d\varepsilon^{-3} \log(\mathrm{e} + dT/\varepsilon)$ and $C_{\mathrm{pc}} = O(1)$, we can apply Corollary 12 to obtain the following small-loss regret bound.

Theorem 15. *For online ℓ_1 regression under the random-order model, Algorithm 1 with $\lambda = H_T^{-4/3}$ runs in polynomial time and achieves $\mathcal{R}_T = \tilde{O}\big(d + d^{1/4}\mathrm{OPT}_T^{3/4}\big)$.*

Proof. Corollary 12 with $q = 3$, $C_1 = C_{\ell_1} d$, $C_2 = dT$, and $C_{\mathrm{pc}} = O(1)$ gives the regret bound. For the time-complexity claim, the number of sampled rows depends on the accuracy parameter. Put $\varepsilon_\star := \varepsilon_{\min}(T + \lambda)$. Comparing its defining objective value with that of the feasible point $\varepsilon = 1$ gives $C_{\ell_1} d\varepsilon_\star^{-3} \leq T + \lambda + C_{\ell_1} d \log(\mathrm{e} + dT)$, and hence $\varepsilon_\star^{-1}$ is polynomially bounded in d and T. Let $c_{\ell_1} \geq 1$ be the universal sample-count coefficient (see the proof of Lemma L.4). The augmented matrix has $d+1$ columns, and the mean sample count $\mu_\varepsilon = c_{\ell_1}(d+1)\varepsilon^{-2} \log(\mathrm{e} + 2(d+1)T/\varepsilon) + 2$ is nonincreasing in ε. Thus, every realized sample count is bounded by $\lceil \mu_{\varepsilon_\star} \rceil$ (see Lemma I.4), which is polynomial in d and T. Together with the deterministic polynomial-time approximate Lewis-weight computation in Appendix L and the assumed polynomial-time exact optimizer, this count bound keeps the sampled weighted instance polynomial in size and makes every round run in polynomial time in the exact-real model. □

Online ℓ_1 regression is a classical research topic [10, 12]. The absolute loss is not self-bounding, and hence the standard self-bounding argument for small-loss regret does not apply (e.g., Orabona [45, Section 4.2]). Our approach obtains a small-loss regret bound under the random-order model by leveraging the average-stable offline algorithm based on Lewis-weight sampling.

7 Further applications and a general idea for achieving small-loss bounds

Dong and Yoshida [19] also obtained fixed-ε, $(1 + \varepsilon)$-approximate regret bounds for online (k, z)-clustering and online low-rank matrix approximation. In Appendix M, we give offline algorithms for these problems that satisfy our assumptions and yield the corresponding small-loss regret bounds.

Having observed a variety of applications, we now discuss a general idea for achieving small-loss regret bounds. Our framework hinges on average-stable offline algorithms with *multiplicative* $(1+\varepsilon)$-approximation guarantees and stability under changes of the accuracy parameter on the same input.

The approximation guarantee and the average-sensitivity bound control the accuracy–stability trade-off through the choice of ε_t (Lemma 8), while parameter stability controls the parameter-change cost (Lemma 10). Once these two parts are controlled, Theorem 11 yields a small-loss regret bound in terms of $\varphi^\star(\mathrm{OPT}_T)$. Algorithmic compression techniques—sparsification, sampling, coresets, and sketching—are natural ingredients for this approach: the resulting offline algorithms provide multiplicative approximation guarantees, and changing the input or the accuracy parameter affects intermediate quantities such as sampling probabilities, sample counts, or reweighting factors, whose total-variation changes can be bounded directly. Concretely, the $(1 \pm \varepsilon)$-cut sparsifier for submodular hypergraphs and Lewis-weight sampling instantiate this mechanism in online submodular function minimization and online ℓ_1 regression, respectively; coreset construction [26] and matrix sketching [15] play analogous roles in online (k, z)-clustering and online low-rank matrix approximation, as detailed in Appendix M. More broadly, sparsifiers for sums of functions such as p-norms and generalized linear models [30, 31] provide promising candidates for further applications once the corresponding stability bounds are verified. This concrete route from offline compression algorithms to small-loss regret guarantees under the random-order model constitutes one conceptual contribution of our work.

8 Conclusion and limitations

Starting from the fixed-ε, $(1+\varepsilon)$-approximate-regret framework of Dong and Yoshida [19], we have established a small-loss regret bound typically of order $\tilde{O}(\varphi^\star(\mathrm{OPT}_T))$ under the random-order model. Our core idea is to allow approximation accuracy ε_t to change over time, which leads to a novel decomposition of the regret into the accuracy–stability trade-off and the parameter-change cost parts. We bound both parts by setting ε_t to optimize the former trade-off part and leveraging stability of offline algorithms with respect to parameter changes, together with the properties of the concave conjugate $\varphi^\star$. Crucially, our framework does not require structural assumptions on loss functions, such as linearity or smoothness, and applies broadly to any problems admitting offline approximation algorithms with bounded average sensitivity and parameter-change stability. We have instantiated our framework for various problems, including online submodular function minimization and online ℓ_1 regression. Beyond these specific applications, we have discussed how our framework can serve as a general recipe that connects widely used algorithmic compression techniques—such as sparsifiers, sampling, coresets, and sketching—with small-loss regret bounds in online learning under the random-order model.

We close with a few remarks on limitations and open questions. First, our results are limited to the random-order model. As a generalization, the *stochastically extended adversarial model* has attracted growing attention [14, 47, 59], and understanding how our average sensitivity-based approach extends to this more general model is an intriguing direction for future work. Second, our method runs an offline algorithm on the data observed so far at each round, and the resulting computational cost can be high. Integrating dynamic or incremental implementations of the required sparsifiers, sketches, or sampling distributions [51, 61] is an important direction for improving practicality. Also, the online (k, z)-clustering instantiation in Appendix M.1 assumes exact evaluation of the optimum on each observed prefix and deterministic exact optimization of each weighted coreset objective; therefore, we do not claim a polynomial-time guarantee for this instantiation. Third, the tightness of our small-loss regret bounds remains open. For online submodular function minimization with $1 \le n \le T$, every learner incurs expected regret $\Omega(\sqrt{nT})$ on some fixed multiset presented in uniformly random order, and hence $\Omega(\sqrt{n\mathrm{OPT}_T})$ regret on that multiset (see Appendix N for details). If we do not care about efficiency, the multiplicative weights update with 2^n experts always achieves $O(\sqrt{n\mathrm{OPT}_T} + n)$. However, focusing on polynomial-time algorithms in the regime of $\mathrm{OPT}_T = o(T)$, exploring improvements over our bound of $\tilde{O}\big(n^3 + n^{3/4}\mathrm{OPT}_T^{3/4}\big)$ is an interesting question for future work. Finally, although we have verified the parameter-change stability requirement in Assumption 9 for all relevant applications, it remains desirable to identify more intuitive and easily checkable sufficient conditions for this assumption. Such an investigation may further clarify what properties of offline algorithms enable small-loss regret bounds through our framework.

Acknowledgments and Disclosure of Funding

Shinsaku Sakaue was supported by JST BOOST Program Japan Grant Number JPMJBY24D1. Yuichi Yoshida was supported by JSPS KAKENHI Grant Number JP24K02903. The authors declare no competing interests.

References

[1] A. Abate, F. Redig, and I. Tkachev. On the effect of perturbation of conditional probabilities in total variation. *Statistics & Probability Letters*, 88:1–8, 2014 (cited on page 23).

[2] J. D. Abernethy, Y. H. Jung, C. Lee, A. McMillan, and A. Tewari. Online learning via the differential privacy lens. In *Advances in Neural Information Processing Systems*, volume 32, pages 8894–8904. Curran Associates, Inc., 2019 (cited on pages 1, 3).

[3] A. Agarwal, A. Krishnamurthy, J. Langford, H. Luo, and R. E. Schapire. Open problem: First-order regret bounds for contextual bandits. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 4–7. PMLR, 2017 (cited on page 2).

[4] A. Agarwal and E. Hazan. New algorithms for repeated play and universal portfolio management. Technical report TR-740-05, Princeton University, 2005 (cited on page 1).

[5] Z. Allen-Zhu, S. Bubeck, and Y. Li. Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 186–194. PMLR, 2018 (cited on page 2).

[6] N. Alon, M. Bun, R. Livni, M. Malliaris, and S. Moran. Private and online learnability are equivalent. *Journal of the ACM*, 69(4), 2022 (cited on page 1).

[7] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007 (cited on page 47).

[8] M. Babaioff, N. Immorlica, D. Kempe, and R. Kleinberg. Matroid secretary problems. *Journal of the ACM*, 65(6), 2018 (cited on pages 1, 3).

[9] M. Bernasconi, A. Celli, R. Colini Baldeschi, F. Fusco, S. Leonardi, and M. Russo. Online learning in the random-order model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 3899–3919. PMLR, 2025 (cited on pages 1, 3).

[10] E. J. Bernstein. Absolute error bounds for learning linear functions online. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 160–163. Association for Computing Machinery, 1992 (cited on page 9).

[11] A. Block and A. Shetty. Small loss bounds for online learning separated function classes: A Gaussian process perspective. *arXiv:2502.10292*, 2025 (cited on page 3).

[12] T. Bylander. Worst-case absolute loss bounds for linear learning algorithms. In *Proceedings of the 14th National Conference on Artificial Intelligence and 9th Conference on Innovative Applications of Artificial Intelligence*, pages 485–490. AAAI Press, 1997 (cited on page 9).

[13] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006 (cited on pages 1, 2, 8, 21).

[14] S. Chen, Y.-J. Zhang, W.-W. Tu, P. Zhao, and L. Zhang. Optimistic online mirror descent for bridging stochastic and adversarial online convex optimization. *Journal of Machine Learning Research*, 25(178):1–62, 2024 (cited on page 10).

[15] M. B. Cohen, C. Musco, and C. Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the 2017 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1758–1777. Society for Industrial and Applied Mathematics, 2017 (cited on pages 10, 59, 63).

[16] M. B. Cohen and R. Peng. ℓ_p row sampling by Lewis weights. *arXiv:1412.0588v1*, 2014. https://arxiv.org/abs/1412.0588v1 (cited on pages 36, 38, 39).

[17] M. B. Cohen and R. Peng. ℓ_p row sampling by Lewis weights. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 183–192. Association for Computing Machinery, 2015 (cited on pages 9, 35, 36, 38, 43).

[18] V. Cohen-Addad, B. Guedj, V. Kanade, and G. Rom. Online k-means clustering. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1126–1134. PMLR, 2021 (cited on page 46).

[19] J. Dong and Y. Yoshida. A batch-to-online transformation under random-order model. In *Advances in Neural Information Processing Systems*, volume 36, pages 56278–56297. Curran Associates, Inc., 2023 (cited on pages 1–5, 7, 9, 10, 18, 20–22, 32, 35, 58).

[20] E. B. Dynkin. Optimal choice of the stopping moment of a Markov process. *Doklady Akademii Nauk SSSR*, 150(2):238–240, 1963 (cited on pages 1, 3).

[21] D. J. Foster and A. Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Advances in Neural Information Processing Systems*, volume 34, pages 18907–18919. Curran Associates, Inc., 2021 (cited on page 2).

[22] D. Garber, G. Korcia, and K. Levy. Online convex optimization in the random order model. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 3387–3396. PMLR, 2020 (cited on pages 1, 3).

[23] A. Gonen, E. Hazan, and S. Moran. Private learning implies online learning: An efficient reduction. In *Advances in Neural Information Processing Systems*, volume 32, pages 8702–8712. Curran Associates, Inc., 2019 (cited on page 1).

[24] A. Gupta and S. Singla. Random-order models. In *Beyond the Worst-Case Analysis of Algorithms*, pages 234–258. Cambridge University Press, 2021 (cited on pages 1, 3).

[25] E. Hazan and S. Kale. Online submodular minimization. *Journal of Machine Learning Research*, 13(93):2903–2922, 2012 (cited on pages 2, 7, 9, 63).

[26] L. Huang and N. K. Vishnoi. Coresets for clustering in Euclidean spaces: Importance sampling is nearly optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1416–1429. Association for Computing Machinery, 2020 (cited on pages 10, 46, 55–57).

[27] S. Ito. Revisiting online submodular minimization: Gap-dependent regret bounds, best of both worlds and adversarial robustness. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9678–9694. PMLR, 2022 (cited on pages 2, 9).

[28] S. Ito, S. Hirahara, T. Soma, and Y. Yoshida. Tight first- and second-order regret bounds for adversarial linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 2028–2038. Curran Associates, Inc., 2020 (cited on page 2).

[29] S. Iwata, L. Fleischer, and S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48:761–777, 4, 2001 (cited on page 31).

[30] A. Jambulapati, J. R. Lee, Y. P. Liu, and A. Sidford. Sparsifying sums of norms. In *Proceedings of the 2023 IEEE 64th Annual Symposium on Foundations of Computer Science*, pages 1953–1962. IEEE, 2023 (cited on page 10).

[31] A. Jambulapati, J. R. Lee, Y. P. Liu, and A. Sidford. Sparsifying generalized linear models. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1665–1675. Association for Computing Machinery, 2024 (cited on page 10).

[32] H. Jiang, Y. T. Lee, Z. Song, and L. Zhang. Convex minimization with integer minima in $\widetilde{O}(n^4)$ time. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 3659–3684. Society for Industrial and Applied Mathematics, 2024 (cited on page 31).

[33] R. M. Karp, U. V. Vazirani, and V. V. Vazirani. An optimal algorithm for on-line bipartite matching. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, pages 352–358. Association for Computing Machinery, 1990 (cited on pages 1, 3).

[34] Y. Kenneth and R. Krauthgamer. Cut sparsification and succinct representation of submodular hypergraphs. In *Proceedings of the 51st International Colloquium on Automata, Languages, and Programming*, volume 297, 97:1–97:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024 (cited on pages 2, 7, 8, 29, 30).

[35] S. Kumabe and Y. Yoshida. Average sensitivity of dynamic programming. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1925–1961. Society for Industrial and Applied Mathematics, 2022 (cited on pages 1, 24).

[36] C.-W. Lee, H. Luo, C.-Y. Wei, and M. Zhang. Bias no more: High-probability data-dependent regret bounds for adversarial bandits and MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 15522–15533. Curran Associates, Inc., 2020 (cited on page 2).

[37] C.-W. Lee, H. Luo, and M. Zhang. A closer look at small-loss bounds for bandits with graph feedback. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125, pages 2516–2564. PMLR, 2020 (cited on page 2).

[38] T. Lindvall. *Lectures on the Coupling Method*. John Wiley & Sons, 1992 (cited on page 23).

[39] T. Lykouris, K. Sridharan, and É. Tardos. Small-loss bounds for online learning with partial information. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pages 979–986. PMLR, 2018 (cited on pages 2, 3).

[40] H. Narayanan and A. Rakhlin. Random walk approach to regret minimization. In *Advances in Neural Information Processing Systems*, volume 23, pages 1777–1785. Curran Associates, Inc., 2010 (cited on page 8).

[41] G. Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of the 28th Conference on Learning Theory*, volume 40, pages 1360–1375. PMLR, 2015 (cited on page 2).

[42] R. Niazadeh, T. Roughgarden, and J. R. Wang. Optimal algorithms for continuous non-monotone submodular and DR-submodular maximization. *Journal of Machine Learning Research*, 21(125):1–31, 2020 (cited on page 2).

[43] J. Nie, W. Kotlowski, and M. K. Warmuth. Online PCA with optimal regret. *Journal of Machine Learning Research*, 17(173):1–49, 2016 (cited on page 58).

[44] J. Olkhovskaya, J. Mayo, T. van Erven, G. Neu, and C.-Y. Wei. First- and second-order bounds for adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems*, volume 36, pages 61625–61644. Curran Associates, Inc., 2023 (cited on page 2).

[45] F. Orabona. Online learning: A modern introduction using convex optimization. *arXiv:1912.13213v10*, 2026. https://arxiv.org/abs/1912.13213v10 (cited on pages 2, 9).

[46] A. Parulekar, A. Parulekar, and E. Price. L1 regression with Lewis weights subsampling. In *Proceedings of the 2021 Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 207, 49:1–49:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021 (cited on pages 2, 9, 35).

[47] S. Sachs, H. Hadiji, T. van Erven, and C. Guzmán. Between stochastic and adversarial online convex optimization: Improved regret bounds via smoothness. In *Advances in Neural Information Processing Systems*, volume 35, pages 691–702. Curran Associates, Inc., 2022 (cited on page 10).

[48] A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80:346–355, 2, 2000 (cited on page 31).

[49] S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2–3):115–142, 2007 (cited on page 1).

[50] U. Sherman, T. Koren, and Y. Mansour. Optimal rates for random order online optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 2097–2108. Curran Associates, Inc., 2021 (cited on pages 1, 3).

[51] T. Soma, K. C. Tung, and Y. Yoshida. Online algorithms for spectral hypergraph sparsification. In *Proceedings of the 25th Conference on Integer Programming and Combinatorial Optimization*, pages 405–417. Springer Nature Switzerland, 2024 (cited on page 10).

[52] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, volume 23, pages 2199–2207. Curran Associates, Inc., 2010 (cited on page 3).

[53] M. Streeter and D. Golovin. An online algorithm for maximizing submodular functions. In *Advances in Neural Information Processing Systems*, volume 21, pages 1577–1584. Curran Associates, Inc., 2008 (cited on page 2).

[54] C.-E. Tsai, Y.-T. Lin, and Y.-H. Li. Data-dependent bounds for online portfolio selection without Lipschitzness and smoothness. In *Advances in Neural Information Processing Systems*, volume 36, pages 62764–62791. Curran Associates, Inc., 2023 (cited on page 3).

[55] N. Varma and Y. Yoshida. Average sensitivity of graph algorithms. In *Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 684–703. Society for Industrial and Applied Mathematics, 2021 (cited on page 1).

[56] N. Varma and Y. Yoshida. Average sensitivity of graph algorithms. *SIAM Journal on Computing*, 52(4):1039–1081, 2023 (cited on pages 1, 3).

[57] G. Wang, Z. Hu, V. Muthukumar, and J. D. Abernethy. Adaptive oracle-efficient online learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23398–23411. Curran Associates, Inc., 2022 (cited on page 3).

[58] K. Wang, K. Zhou, R. Wu, N. Kallus, and W. Sun. The benefits of being distributional: Small-loss bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 2275–2312. Curran Associates, Inc., 2023 (cited on page 2).

[59] Y. Wang, S. Chen, W. Jiang, W. Yang, Y. Wan, and L. Zhang. Online composite optimization between stochastic and adversarial environments. In *Advances in Neural Information Processing Systems*, volume 37, pages 94808–94850. Curran Associates, Inc., 2024 (cited on page 10).

[60] M. K. Warmuth and D. Kuzmin. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9(75):2287–2320, 2008 (cited on page 58).

[61] D. P. Woodruff and T. Yasuda. Online Lewis weight sampling. *ACM Transactions on Algorithms*, 21, 4, 2025 (cited on page 10).

[62] W. Yang, W. Jiang, Y. Wang, P. Yang, Y. Hu, and L. Zhang. Small-loss adaptive regret for online convex optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 56156–56195. PMLR, 2024 (cited on page 3).

[63] Y. Yoshida and S. Ito. Average sensitivity of Euclidean k-clustering. In *Advances in Neural Information Processing Systems*, volume 35, pages 32487–32498. Curran Associates, Inc., 2022 (cited on pages 1, 3, 7, 18, 32, 46, 47, 50).

[64] P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1–52, 2024 (cited on page 3).

A Proof of Lemma 1

Lemma 1 (Bound on sum of average optimal values)**.** *It holds that* $\mathbb{E}\big[\sum_{t=1}^{T} \mathrm{OPT}_t/t\big] \le \mathrm{OPT}_T$.

Proof. Let $\theta^\star \in \Theta$ attain OPT_T on the fixed input multiset X_T. Recall that the arrival sequence $x_1, \dots, x_T$ is a uniformly random ordering of its labeled occurrences. For every $t \in [T]$, we have

$$\mathrm{OPT}_t = \min_{\theta\in\Theta} \sum_{i=1}^{t} \ell(\theta, x_i) \le \sum_{i=1}^{t} \ell(\theta^\star, x_i).$$

Each labeled occurrence in X_T appears in the first t positions with probability t/T. Taking expectation over the random order therefore gives

$$\mathbb{E}[\mathrm{OPT}_t] \le \frac{t}{T} \sum_{x\in X_T} \ell(\theta^\star, x) = \frac{t}{T}\mathrm{OPT}_T.$$

Dividing by t and summing for $t = 1, \dots, T$ yields the claim. □

B Proof of Proposition 5

The following lemma is central to the derivation of the regret decomposition in Proposition 5.

Lemma B.1 (Single-step bound)**.** *Let $t \in \{1, \dots, T-1\}$ and $\theta_{t+1} \in \Theta$ denote the output of $\mathcal{A}_{\varepsilon_t}$ in Algorithm 1. Suppose $\varepsilon_t = \varepsilon_{X_t}$ is specified by a deterministic measurable rule $Y \mapsto \varepsilon_Y$ that depends only on the multiset Y. Then, it holds that*

$$\begin{aligned}&\mathbb{E}[\ell(\theta_{t+1}, x_{t+1})]\\ &\le \mathbb{E}\left[\frac{1+\varepsilon_t}{t}\,\mathrm{OPT}_t + \frac{\varphi(\varepsilon_t)}{t} + \frac{2}{t}\sum_{x\in X_t} \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t \setminus x)\Big)\right] + O(1/T).\end{aligned}$$

Proof. Fix $t \in \{1, \dots, T-1\}$ and let $X_t = \{x_1, \dots, x_t\}$. Taking expectation over the randomness of the offline algorithms and the random order of $\{x_t\}_{t=1}^{T}$, we start from

$$\mathbb{E}\left[\ell(\theta_{t+1}, x_{t+1}) - \frac{1}{t}\ell(\theta_{t+1}, X_t)\right] = \mathbb{E}\left[\frac{1}{t}\sum_{x\in X_t}\big(\ell(\mathcal{A}_{\varepsilon_{X_t}}(X_t), x_{t+1}) - \ell(\mathcal{A}_{\varepsilon_{X_t}}(X_t), x)\big)\right],$$

which measures how far the fresh loss $\ell(\theta_{t+1}, x_{t+1})$ can drift from the empirical average over X_t. For $x \in X_t$, add and subtract the corresponding loss term obtained by running the algorithm on $X_t \setminus x$ with the parameter recomputed from $X_t \setminus x$, namely $\ell(\mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t \setminus x), x)$, inside each summand on the right-hand side to obtain

$$\begin{aligned}\mathbb{E}\left[\ell(\theta_{t+1}, x_{t+1}) - \frac{1}{t}\ell(\theta_{t+1}, X_t)\right] &= \mathbb{E}\left[\frac{1}{t}\sum_{x\in X_t}\Big(\ell(\mathcal{A}_{\varepsilon_{X_t}}(X_t), x_{t+1}) - \ell(\mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t \setminus x), x)\Big)\right]\\ &\quad + \mathbb{E}\left[\frac{1}{t}\sum_{x\in X_t}\Big(\ell(\mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t \setminus x), x) - \ell(\mathcal{A}_{\varepsilon_{X_t}}(X_t), x)\Big)\right].\end{aligned}$$

Here, the element removed uniformly from X_t has the same distribution as x_{t+1} under the random-order model. Formally, sample an index $I \sim \mathrm{Unif}(\{1, \dots, t\})$ and set $x = x_I$ and $U = X_t \setminus x_I$. Conditioned on U, the two held-out points x_I and x_{t+1} are exchangeable under a uniformly random permutation of X_{t+1} (they are exactly the two elements of $X_{t+1} \setminus U$). Hence, recalling that ε_U is a function only of the multiset U, we have $\mathbb{E}[\ell(\mathcal{A}_{\varepsilon_U}(U), x_I) \mid U] = \mathbb{E}[\ell(\mathcal{A}_{\varepsilon_U}(U), x_{t+1}) \mid U]$, which allows us to replace x in the first expectation by x_{t+1}.[7] Therefore, we obtain

$$\begin{aligned}\mathbb{E}\left[\ell(\theta_{t+1}, x_{t+1}) - \frac{1}{t}\ell(\theta_{t+1}, X_t)\right] &= \mathbb{E}\left[\frac{1}{t}\sum_{x\in X_t}\Big(\ell(\mathcal{A}_{\varepsilon_{X_t}}(X_t), x_{t+1}) - \ell(\mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t \setminus x), x_{t+1})\Big)\right]\\ &\quad + \mathbb{E}\left[\frac{1}{t}\sum_{x\in X_t}\Big(\ell(\mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t \setminus x), x) - \ell(\mathcal{A}_{\varepsilon_{X_t}}(X_t), x)\Big)\right].\end{aligned} \tag{3}$$

[7] The point of the argument is to use the parameter recomputed from the reduced dataset when invoking exchangeability, and then bound the difference from the original parameter through the separate parameter-change term.

Here, the range of $\ell(\cdot, x')$ is $[0,1]$ for any $x' \in \mathcal{X}$, which implies

$$\left|\mathbb{E}_{\mathcal{A}}\left[\ell(\mathcal{A}_{\varepsilon_{X_t}}(X_t), x') - \ell(\mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x), x')\right]\right| \le \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big),$$

where $\mathbb{E}_{\mathcal{A}}$ denotes expectation over the randomness of the offline algorithms, with X_t, x, and x' held fixed. Therefore, the above two components are bounded above by $\mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big)$ in expectation. We next split this total variation distance by the triangle inequality into a fixed-ε deletion part and a term comparing two parameters on the same reduced dataset:

$$\begin{aligned}
&\mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big) \\
&\le \mathrm{TV}\big(\mathcal{A}_{\varepsilon_{X_t}}(X_t), \mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x)\big) + \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big).
\end{aligned}$$

Consequently, using the fixed-ε_t average sensitivity bound in Assumption 4, we obtain

$$\mathbb{E}[\ell(\theta_{t+1}, x_{t+1})] \le \mathbb{E}\left[\frac{1}{t}\ell(\theta_{t+1}, X_t) + \frac{\varphi(\varepsilon_t)}{t} + \frac{2}{t}\sum_{x \in X_t} \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big)\right],$$

where the factor 2 comes from the two terms on the right-hand side in (3). By $(1+\varepsilon_t)$-approximation of $\mathcal{A}_{\varepsilon_t}$ in Assumption 4, for any $X_t \in \mathcal{X}^t$, we have

$$\mathbb{E}_{\mathcal{A}_{\varepsilon_t}}[\ell(\theta_{t+1}, X_t)] = \mathbb{E}_{\mathcal{A}_{\varepsilon_t}}[\ell(\mathcal{A}_{\varepsilon_t}(X_t), X_t)] \le (1+\varepsilon_t)\mathrm{OPT}_t + O(t/T).$$

Combining the above two inequalities and taking expectation over the random order of $\{x_t\}_{t=1}^T$, we obtain

$$\begin{aligned}
&\mathbb{E}[\ell(\theta_{t+1}, x_{t+1})] \\
&\le \mathbb{E}\left[\frac{1+\varepsilon_t}{t}\,\mathrm{OPT}_t + \frac{\varphi(\varepsilon_t)}{t} + \frac{2}{t}\sum_{x \in X_t} \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big)\right] + O(1/T),
\end{aligned}$$

completing the proof. □

By using the above single-step bound and the bound on the sum of average optimal values in Lemma 1, we can now prove the regret decomposition in Proposition 5.

Proposition 5 (Regret decomposition)**.** *Suppose that the parameters ε_t used by Algorithm 1 are specified by a measurable rule $Y \mapsto \varepsilon_Y$ for any finite multiset Y, so that $\varepsilon_t = \varepsilon_{X_t}$ holds for every t; let $\varepsilon_\emptyset \in (0,1]$ be an arbitrary fixed value. For $\theta_1, \dots, \theta_T \in \Theta$ obtained by Algorithm 1, the regret $\mathcal{R}_T$ is upper bounded by the following two sums up to an additive constant:*[8]

$$\mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\left(\frac{\varepsilon_t}{t}\mathrm{OPT}_t + \frac{\varphi(\varepsilon_t)}{t}\right)}_{\text{accuracy–stability trade-off}}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\frac{2}{t}\sum_{x \in X_t} \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big)}_{\text{parameter-change cost}}\right] + O(1).$$

Proof. Summing both sides of Lemma B.1 for $t = 1, \dots, T-1$ and using $\mathbb{E}\left[\frac{1+\varepsilon_T}{T}\,\mathrm{OPT}_T + \frac{\varphi(\varepsilon_T)}{T}\right] \ge 0$ yield the following inequality up to an additive $O(1)$ term:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T-1} \ell(\theta_{t+1}, x_{t+1})\right] \\
&\le \sum_{t=1}^{T-1}\left(\mathbb{E}\left[\frac{1+\varepsilon_t}{t}\,\mathrm{OPT}_t + \frac{\varphi(\varepsilon_t)}{t}\right] + \mathbb{E}\left[\frac{2}{t}\sum_{x \in X_t} \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big)\right]\right) \\
&\le \sum_{t=1}^{T}\left(\mathbb{E}\left[\frac{1+\varepsilon_t}{t}\,\mathrm{OPT}_t + \frac{\varphi(\varepsilon_t)}{t}\right] + \mathbb{E}\left[\frac{2}{t}\sum_{x \in X_t} \mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t \setminus x), \mathcal{A}_{\varepsilon_{X_t \setminus x}}(X_t \setminus x)\Big)\right]\right).
\end{aligned}$$

[8] Terms involving the empty input contribute only an additive $O(1)$.

The left-hand side is bounded below by $\mathbb{E}\left[\sum_{t=1}^{T} \ell(\theta_t, x_t)\right] - 1$, while Lemma 1 implies that the right-hand side is bounded above by

$$\mathrm{OPT}_T + \sum_{t=1}^{T}\left(\mathbb{E}\left[\frac{\varepsilon_t}{t}\,\mathrm{OPT}_t + \frac{\varphi(\varepsilon_t)}{t}\right] + \mathbb{E}\left[\frac{2}{t}\sum_{x\in X_t}\mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t\setminus x), \mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t\setminus x)\Big)\right]\right).$$

Rearranging the terms completes the proof. □

C Proof of Lemma 7

Lemma 7. *The function $\varphi^\star$ is finite-valued, non-decreasing, and concave on $\mathbb{R}_{\geq 0}$. The selected minimizer $\varepsilon_{\min}\colon \mathbb{R}_{>0} \to (0,1]$ is non-increasing and, for every $u > 0$, is a supergradient of $\varphi^\star$ at u, i.e., $\varepsilon_{\min}(u)(u - v) \leq \varphi^\star(u) - \varphi^\star(v)$ for every $v \geq 0$.*

Proof. Because φ is proper, fix $\bar{\varepsilon} \in \mathrm{dom}(\varphi)$. For every $u \geq 0$, we have

$$0 \leq \varphi^\star(u) \leq u\bar{\varepsilon} + \varphi(\bar{\varepsilon}) < +\infty,$$

and hence $\varphi^\star$ is finite-valued on $\mathbb{R}_{\geq 0}$. It is also non-decreasing: if $0 \leq u_1 < u_2$, then we have

$$u_1\varepsilon + \varphi(\varepsilon) \leq u_2\varepsilon + \varphi(\varepsilon) \qquad \text{for every } \varepsilon \in \mathrm{dom}(\varphi),$$

and taking infima over ε gives $\varphi^\star(u_1) \leq \varphi^\star(u_2)$. The concavity of $\varphi^\star$ follows because it is the pointwise infimum of the affine functions $u \mapsto u\varepsilon + \varphi(\varepsilon)$ indexed by $\varepsilon \in \mathrm{dom}(\varphi)$. The selected minimizer satisfies $\varepsilon_{\min}(u) \in (0,1]$ as the effective domain of φ is contained in $(0,1]$, and it is non-increasing in u: if $0 < u_1 < u_2$ and $\varepsilon_i = \varepsilon_{\min}(u_i)$, then adding the two optimality inequalities,

$$u_1\varepsilon_1 + \varphi(\varepsilon_1) \leq u_1\varepsilon_2 + \varphi(\varepsilon_2) \qquad \text{and} \qquad u_2\varepsilon_2 + \varphi(\varepsilon_2) \leq u_2\varepsilon_1 + \varphi(\varepsilon_1),$$

yields $0 \leq (u_2 - u_1)(\varepsilon_1 - \varepsilon_2)$, hence $\varepsilon_1 \geq \varepsilon_2$. Finally, we verify the supergradient claim. Fix $u > 0$ and write $\varepsilon_u = \varepsilon_{\min}(u)$. By the optimality of ε_u, for any $v \geq 0$, we have

$$\varphi^\star(v) = \inf_{\varepsilon\in(0,1]}\{v\varepsilon + \varphi(\varepsilon)\} \leq v\varepsilon_u + \varphi(\varepsilon_u) = \varphi^\star(u) + \varepsilon_u(v - u).$$

This is exactly the supergradient inequality for the concave function $\varphi^\star$ at u. □

D Proof of Lemma 10

Lemma 10 (Controlling parameter-change cost)**.** *Let $\mathrm{OPT}_0 = 0$ and assume Assumption 9. Then, the parameter-change cost part in Proposition 5 is bounded as follows:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{2}{t}\sum_{x\in X_t}\mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t\setminus x), \mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t\setminus x)\Big)\right] \leq C_{\mathrm{pc}}\mathbb{E}\left[\sum_{t=1}^{T}\varepsilon_t(\mathrm{OPT}_t - \mathrm{OPT}_{t-1})\right] + O(1).$$

Proof. The $t = 1$ summand is at most 2 and is absorbed into the $O(1)$ term. For $t \geq 2$, Assumption 9 applied to X_t gives

$$\frac{2}{t}\sum_{x\in X_t}\mathrm{TV}\Big(\mathcal{A}_{\varepsilon_{X_t}}(X_t\setminus x), \mathcal{A}_{\varepsilon_{X_t\setminus x}}(X_t\setminus x)\Big) \leq \frac{C_{\mathrm{pc}}\varepsilon_t}{t}\sum_{x\in X_t}(\mathrm{OPT}_t - \mathrm{OPT}(X_t\setminus x)).$$

Conditioned on X_t, the prefix X_{t-1} is distributed as $X_t \setminus x$ with $x \in X_t$ being a uniformly chosen occurrence, and ε_t depends only on X_t. Thus, we have $\mathbb{E}[\,\mathrm{OPT}_{t-1} \mid X_t\,] = \frac{1}{t}\sum_{x\in X_t}\mathrm{OPT}(X_t\setminus x)$, and hence taking expectation over X_t gives

$$\mathbb{E}\left[\frac{C_{\mathrm{pc}}\varepsilon_t}{t}\sum_{x\in X_t}(\mathrm{OPT}_t - \mathrm{OPT}(X_t\setminus x))\right] = C_{\mathrm{pc}}\mathbb{E}[\varepsilon_t(\mathrm{OPT}_t - \mathrm{OPT}_{t-1})].$$

Summing over t completes the proof. □

E Proof of Corollary 12

Corollary 12. *Let $q > 0$ be a constant. Assume $\varphi(\varepsilon) = C_1\varepsilon^{-q}\log(\mathrm{e} + C_2/\varepsilon)$ for $\varepsilon \in (0,1]$ and $\varphi(\varepsilon) = +\infty$ for $\varepsilon \notin (0,1]$ for some $C_1 \geq 1$ and $C_2 \geq 0$. Then, for any $\lambda \in (0, H_T^{-(q+1)/q}]$ such that Assumption 9 holds, Theorem 11 yields $\mathcal{R}_T = \tilde{O}((1+C_{\mathrm{pc}})(C_1 + C_1^{1/(q+1)}\mathrm{OPT}_T^{q/(q+1)}))$.*

Proof. Recall that $\varphi(\varepsilon) = C_1\varepsilon^{-q}\log(\mathrm{e} + C_2/\varepsilon)$ on $(0,1]$ and $\varphi(\varepsilon) = +\infty$ outside $(0,1]$; this choice of φ satisfies the conditions in Assumption 4. Let $L_0 = \log(\mathrm{e} + C_2)$. For $u > 0$, set $\widehat{\varepsilon}(u) = \min\{1, (qC_1/u)^{1/(q+1)}\}$. Substituting this feasible value into $u\varepsilon + \varphi(\varepsilon)$ gives

$$\varphi^\star(u) \leq C_1 L_0 + q^{-\frac{q}{q+1}} C_1^{\frac{1}{q+1}} u^{\frac{q}{q+1}} \left(q + \log\left(\mathrm{e} + C_2 q^{-\frac{1}{q+1}} C_1^{-\frac{1}{q+1}} u^{\frac{1}{q+1}}\right)\right).$$

Substituting $u = \frac{\mathrm{OPT}_T}{H_T} + \lambda$ into the above yields the following upper bound on $\varphi^\star(u)$:

$$C_1 L_0 + q^{-\frac{q}{q+1}} C_1^{\frac{1}{q+1}} \left(\frac{\mathrm{OPT}_T}{H_T} + \lambda\right)^{\frac{q}{q+1}} \left(q + \log\left(\mathrm{e} + C_2 q^{-\frac{1}{q+1}} C_1^{-\frac{1}{q+1}} \left(\frac{\mathrm{OPT}_T}{H_T} + \lambda\right)^{\frac{1}{q+1}}\right)\right).$$

By monotonicity and subadditivity of $u \mapsto u^\alpha$ for $\alpha \in (0,1)$, and $\lambda \leq H_T^{-(q+1)/q} \leq 1$, we have

$$\left(\frac{\mathrm{OPT}_T}{H_T} + \lambda\right)^{\frac{q}{q+1}} \leq \left(\frac{\mathrm{OPT}_T}{H_T}\right)^{\frac{q}{q+1}} + H_T^{-1} \quad \text{and} \quad \left(\frac{\mathrm{OPT}_T}{H_T} + \lambda\right)^{\frac{1}{q+1}} \leq \left(1 + \frac{\mathrm{OPT}_T}{H_T}\right)^{\frac{1}{q+1}}.$$

Therefore, we obtain the following upper bound on $H_T\varphi^\star(u)$:

$$C_1 H_T L_0 + q^{-\frac{q}{q+1}} C_1^{\frac{1}{q+1}} \left(1 + \mathrm{OPT}_T^{\frac{q}{q+1}} H_T^{\frac{1}{q+1}}\right) \left(q + \log\left(\mathrm{e} + C_2 \left(\frac{1 + \mathrm{OPT}_T/H_T}{qC_1}\right)^{\frac{1}{q+1}}\right)\right).$$

It remains to bound the second term in Theorem 11. Since $\varphi^\star$ is non-negative and non-decreasing, it holds that

$$\varphi^\star(\mathrm{OPT}_T + \lambda) - \varphi^\star(\lambda) \leq \varphi^\star(\mathrm{OPT}_T + \lambda).$$

Applying the same estimate for $\varphi^\star(u)$ with $u = \mathrm{OPT}_T + \lambda$ and using $\lambda \leq 1$ gives the following upper bound on $\varphi^\star(\mathrm{OPT}_T + \lambda)$:

$$C_1 L_0 + q^{-\frac{q}{q+1}} C_1^{\frac{1}{q+1}} \left(1 + \mathrm{OPT}_T^{\frac{q}{q+1}}\right) \left(q + \log\left(\mathrm{e} + C_2 \left(\frac{1 + \mathrm{OPT}_T}{qC_1}\right)^{\frac{1}{q+1}}\right)\right).$$

Using these two bounds in Theorem 11 and hiding logarithmic factors in $\tilde{O}$ completes the proof. □

F Small-loss inconsistency bound

We show that the parameter selection rule in (1) also yields a small-loss bound on inconsistency. For $t \in [T-1]$, the indicator $\mathbf{1}\{\theta_t \neq \theta_{t+1}\}$ equals 1 when the two outputs differ and 0 otherwise. We measure inconsistency following Dong and Yoshida [19] by counting changes in the sequence of outputs played by the learner:

$$\mathrm{Inc}_T := \mathbb{E}_{\mathsf{Alg},\{x_t\}}\left[\sum_{t=1}^{T-1} \mathbf{1}\{\theta_t \neq \theta_{t+1}\}\right].$$

We first state the probability-transformation requirement. For two output laws P, Q on Θ, a probability transformation from P to Q means a randomized map which, given an input distributed as P, outputs a random variable distributed as Q; its change probability is the probability that the input and output values differ. The fixed-parameter deletion transformation below is the same probability transportation requirement used by Dong and Yoshida [19], following Yoshida and Ito [63, Definition 4.1 and Lemma 4.5]. The new ingredient in our adaptive setting is the same-input parameter-change transformation, which is needed only because ε_X and $\varepsilon_{X\setminus x}$ may differ.

Assumption F.1 (Probability transformations)**.** Fix $\lambda > 0$ and the rule $\varepsilon_Y = \varepsilon_{\min}(\mathrm{OPT}(Y) + \lambda)$ used in (1). For every multiset X with $|X| \geq 2$ and every occurrence $x \in X$, assume that the following two probability transformations are available.

1. There is a same-input parameter-change transformation $K^{\mathrm{pc}}_{X,x}$ such that, if $\theta^- \sim \mathcal{A}_{\varepsilon_{X\setminus x}}(X \setminus x)$ and θ^0 is the output of $K^{\mathrm{pc}}_{X,x}$ on input θ^-, then it holds that
$$\theta^0 \sim \mathcal{A}_{\varepsilon_X}(X \setminus x) \qquad \text{and} \qquad \Pr\big[\theta^0 \neq \theta^-\big] \leq \mathrm{TV}\big(\mathcal{A}_{\varepsilon_{X\setminus x}}(X \setminus x), \mathcal{A}_{\varepsilon_X}(X \setminus x)\big).$$

2. There is a fixed-parameter deletion transformation $K^{\mathrm{del}}_{X,x}$ such that, if $\theta^0 \sim \mathcal{A}_{\varepsilon_X}(X \setminus x)$ and θ^+ is the output of $K^{\mathrm{del}}_{X,x}$ on input θ^0, then it holds that
$$\theta^+ \sim \mathcal{A}_{\varepsilon_X}(X) \qquad \text{and} \qquad \Pr\big[\theta^+ \neq \theta^0\big] \leq \mathrm{TV}(\mathcal{A}_{\varepsilon_X}(X \setminus x), \mathcal{A}_{\varepsilon_X}(X)).$$

The theorem below is stated for the coupled implementation: after the first prefix, each new output is obtained from the previous one by applying the transformations in Assumption F.1. This preserves the same marginal law $\mathcal{A}_{\varepsilon_t}(X_t)$ at every prefix, and thus the regret analysis in Theorem 11 is unchanged.

Theorem F.2 (Small-loss inconsistency bound)**.** *Under the random-order model and Assumption 4, fix $\lambda > 0$ and suppose that Assumption 9 and Assumption F.1 hold for this λ. Then, the coupled implementation of Algorithm 1 with ε_t given by (1) satisfies*
$$\mathrm{Inc}_T \leq \frac{H_T}{2}\varphi^\star\left(\frac{\mathrm{OPT}_T}{H_T} + \lambda\right) + \frac{C_{\mathrm{pc}}}{2}(\varphi^\star(\mathrm{OPT}_T + \lambda) - \varphi^\star(\lambda)) + O(1).$$

Proof. We first provide the marginal invariant of the coupled implementation. Let $\mathcal{H}_s = (x_1, \ldots, x_s)$ denote the ordered data history up to time s. For every $s \in [T]$, we have
$$\theta_{s+1} \mid \mathcal{H}_s \sim \mathcal{A}_{\varepsilon_{X_s}}(X_s). \tag{4}$$
Indeed, for $s = 1$, this holds by the initialization step after observing x_1. Suppose that the invariant holds at time $t-1$, namely
$$\theta_t \mid \mathcal{H}_{t-1} \sim \mathcal{A}_{\varepsilon_{X_{t-1}}}(X_{t-1}).$$
After observing x_t, write $X = X_t$ and let the newly added occurrence be $x \in X$, so that $X_{t-1} = X \setminus x$. By first applying $K^{\mathrm{pc}}_{X,x}$ and then $K^{\mathrm{del}}_{X,x}$, the output θ_{t+1} has marginal law
$$\theta_{t+1} \mid \mathcal{H}_t \sim \mathcal{A}_{\varepsilon_X}(X) = \mathcal{A}_{\varepsilon_{X_t}}(X_t).$$
Thus, the coupled implementation preserves the same one-prefix marginal law as the independent rerun version of Algorithm 1.

If $T \leq 2$, then $\mathrm{Inc}_T \leq 1$, and the claim follows from the additive constant. Hence assume $T \geq 3$. The change indexed by $t = 1$ in the definition of Inc_T is at most one. We now control the remaining changes. Fix $t \in \{2, \ldots, T-1\}$ and condition on $X_t = X$. Under the random-order model, the previous prefix X_{t-1} is distributed as $X \setminus x$ with x chosen uniformly from the multiset X. For each such occurrence x, the invariant (4) implies
$$\theta_t \mid \{X_t = X,\ X_{t-1} = X \setminus x\} \sim \mathcal{A}_{\varepsilon_{X\setminus x}}(X \setminus x).$$
Indeed, (4) holds conditional on the full ordered history $\mathcal{H}_{t-1}$, and the algorithmic randomness used to generate θ_t is independent of the next unrevealed occurrence once $\mathcal{H}_{t-1}$ is fixed. For this fixed X and x, the union bound and Assumption F.1 give
$$\begin{aligned}&\Pr[\theta_t \neq \theta_{t+1} \mid X_t = X,\ X_{t-1} = X \setminus x]\\ &\leq \mathrm{TV}\big(\mathcal{A}_{\varepsilon_{X\setminus x}}(X \setminus x), \mathcal{A}_{\varepsilon_X}(X \setminus x)\big) + \mathrm{TV}(\mathcal{A}_{\varepsilon_X}(X \setminus x), \mathcal{A}_{\varepsilon_X}(X)).\end{aligned}$$
Averaging over the uniformly deleted occurrence $x \in X$, using the symmetry of total variation, and then applying Assumption 4 with the fixed parameter ε_X, we obtain
$$\begin{aligned}&\Pr[\theta_t \neq \theta_{t+1} \mid X_t = X]\\ &\leq \frac{1}{|X|}\sum_{x\in X}\mathrm{TV}(\mathcal{A}_{\varepsilon_X}(X), \mathcal{A}_{\varepsilon_X}(X \setminus x)) + \frac{1}{|X|}\sum_{x\in X}\mathrm{TV}\big(\mathcal{A}_{\varepsilon_X}(X \setminus x), \mathcal{A}_{\varepsilon_{X\setminus x}}(X \setminus x)\big)\\ &\leq \frac{\varphi(\varepsilon_X)}{2|X|} + \frac{1}{|X|}\sum_{x\in X}\mathrm{TV}\big(\mathcal{A}_{\varepsilon_X}(X \setminus x), \mathcal{A}_{\varepsilon_{X\setminus x}}(X \setminus x)\big).\end{aligned}$$

Applying Assumption 9 to the second term yields, for $X = X_t$,

$$\Pr[\theta_t \neq \theta_{t+1} \mid X_t] \leq \frac{\varphi(\varepsilon_t)}{2t} + \frac{C_{\mathrm{pc}}\varepsilon_t}{2t} \sum_{x \in X_t} (\mathrm{OPT}_t - \mathrm{OPT}(X_t \setminus x)).$$

Taking expectation and using that, conditioned on X_t, the prefix X_{t-1} is distributed as $X_t \setminus x$ for a uniformly chosen occurrence $x \in X_t$, we get

$$\mathbb{E}\left[\frac{C_{\mathrm{pc}}\varepsilon_t}{2t} \sum_{x \in X_t} (\mathrm{OPT}_t - \mathrm{OPT}(X_t \setminus x))\right] = \frac{C_{\mathrm{pc}}}{2}\mathbb{E}[\varepsilon_t(\mathrm{OPT}_t - \mathrm{OPT}_{t-1})].$$

Summing over $t = 2, \ldots, T-1$, adding the possible initial change, and extending the non-negative sums to $t = 1, \ldots, T$ give

$$\mathrm{Inc}_T \leq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \frac{\varphi(\varepsilon_t)}{t}\right] + \frac{C_{\mathrm{pc}}}{2}\mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t(\mathrm{OPT}_t - \mathrm{OPT}_{t-1})\right] + O(1),$$

where $\mathrm{OPT}_0 = 0$.

The rest of the proof bounds the two sums as in the proof of Theorem 11. Since $\mathrm{OPT}_t\varepsilon_t/t \geq 0$, Lemma 8 implies

$$\sum_{t=1}^{T} \frac{\varphi(\varepsilon_t)}{t} \leq \sum_{t=1}^{T}\left(\frac{\mathrm{OPT}_t}{t}\varepsilon_t + \frac{\varphi(\varepsilon_t)}{t}\right) \leq H_T\varphi^\star\left(\frac{\sum_{t=1}^{T}\mathrm{OPT}_t/t}{H_T} + \lambda\right).$$

Taking expectation, applying Jensen's inequality to the concave function $\varphi^\star$, and then using Lemma 1 and the monotonicity of $\varphi^\star$, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{\varphi(\varepsilon_t)}{t}\right] \leq H_T\varphi^\star\left(\frac{\mathrm{OPT}_T}{H_T} + \lambda\right).$$

For the second sum, set $a_t = \mathrm{OPT}_t + \lambda$ and $a_0 = \lambda$. Since the losses are non-negative, OPT_t is non-decreasing along every realized prefix sequence. Hence, by the supergradient property of $\varepsilon_{\min}$ in Lemma 7, we obtain

$$\varepsilon_t(\mathrm{OPT}_t - \mathrm{OPT}_{t-1}) = \varepsilon_{\min}(a_t)(a_t - a_{t-1}) \leq \varphi^\star(a_t) - \varphi^\star(a_{t-1}).$$

Telescoping over $t = 1, \ldots, T$ gives

$$\sum_{t=1}^{T} \varepsilon_t(\mathrm{OPT}_t - \mathrm{OPT}_{t-1}) \leq \varphi^\star(\mathrm{OPT}_T + \lambda) - \varphi^\star(\lambda).$$

Taking expectation of this and combining the two bounds complete the proof. □

This theorem, instantiated with the function φ in Corollary 12, specializes to the following small-loss inconsistency bound, which recovers the fixed-ε bound of Dong and Yoshida [19].

Corollary F.3. *Let $q > 0$ be a constant. Assume $\varphi(\varepsilon) = C_1\varepsilon^{-q}\log(\mathrm{e} + C_2/\varepsilon)$ for $\varepsilon \in (0, 1]$ and $\varphi(\varepsilon) = +\infty$ for $\varepsilon \notin (0, 1]$, where $C_1 \geq 1$ and $C_2 \geq 0$. Under the assumptions of Theorem F.2, for any $\lambda \in (0, H_T^{-(q+1)/q}]$, it holds that*

$$\mathrm{Inc}_T = \tilde{O}\Big((1 + C_{\mathrm{pc}})\Big(C_1 + C_1^{1/(q+1)}\mathrm{OPT}_T^{q/(q+1)}\Big)\Big).$$

Moreover, let $\bar{\varepsilon} \coloneqq \varepsilon_{\min}(\mathrm{OPT}_T + \lambda)$. Then, for any fixed $\varepsilon \in (0, \bar{\varepsilon}]$, the same bound also implies

$$\mathrm{Inc}_T = O((H_T + C_{\mathrm{pc}})\varphi(\varepsilon)).$$

Thus, at the hindsight scale $\bar{\varepsilon}$ (or the known scale $\varepsilon_{\min}(T + \lambda)$ due to the monotonicity of $\varepsilon_{\min}$ and $\ell(\cdot, x_t) \in [0, 1]$) and under $C_{\mathrm{pc}} = O(1)$, the coupled implementation recovers the fixed-ε inconsistency bound of Dong and Yoshida [19].

Proof. The first bound follows from the same estimate on $\varphi^\star$ as in Corollary 12. Indeed, the bound in Theorem F.2 consists of

$$\frac{H_T}{2}\varphi^\star\left(\frac{\mathrm{OPT}_T}{H_T}+\lambda\right) \qquad \text{and} \qquad \frac{C_{\mathrm{pc}}}{2}(\varphi^\star(\mathrm{OPT}_T+\lambda)-\varphi^\star(\lambda)),$$

which are the same two quantities appearing in Theorem 11, up to absolute constants. Therefore, applying the proof of Corollary 12 gives

$$\mathrm{Inc}_T = \tilde{O}\Big((1+C_{\mathrm{pc}})\Big(C_1 + C_1^{1/(q+1)}\mathrm{OPT}_T^{q/(q+1)}\Big)\Big).$$

We next prove the fixed-ε comparison. Since the losses are non-negative, $\mathrm{OPT}_t \le \mathrm{OPT}_T$ holds for every prefix. Hence, by the monotonicity of $\varepsilon_{\min}$ in Lemma 7, it holds that

$$\varepsilon_t = \varepsilon_{\min}(\mathrm{OPT}_t+\lambda) \ge \varepsilon_{\min}(\mathrm{OPT}_T+\lambda) = \bar{\varepsilon} \ge \varepsilon.$$

The function φ is decreasing on $(0,1]$, hence $\varphi(\varepsilon_t) \le \varphi(\varepsilon)$ for every t. Therefore, the fixed-parameter deletion part in the proof of Theorem F.2 is bounded directly as

$$\frac{1}{2}\sum_{t=1}^{T}\frac{\varphi(\varepsilon_t)}{t} \le \frac{H_T}{2}\varphi(\varepsilon).$$

It remains to control the parameter-change part in the proof of Theorem F.2 at the same scale. Put $u = \mathrm{OPT}_T + \lambda$ and $\bar{\varepsilon} = \varepsilon_{\min}(u)$. By optimality, we have $\varphi^\star(u) = u\bar{\varepsilon} + \varphi(\bar{\varepsilon})$. We claim

$$u\bar{\varepsilon} = O_q(\varphi(\bar{\varepsilon})),$$

where the hidden constant depends only on q. Indeed, by comparing the optimal value at $\bar{\varepsilon}$ with the feasible point $\bar{\varepsilon}/2$, we have $u\bar{\varepsilon}+\varphi(\bar{\varepsilon}) \le \frac{u\bar{\varepsilon}}{2}+\varphi(\bar{\varepsilon}/2)$. Thus, we obtain $\frac{u\bar{\varepsilon}}{2} \le \varphi(\bar{\varepsilon}/2)-\varphi(\bar{\varepsilon}) \le \varphi(\bar{\varepsilon}/2)$. For the specific form of φ, we have

$$\varphi(\bar{\varepsilon}/2) = 2^q C_1 \bar{\varepsilon}^{-q}\log(\mathrm{e}+2C_2/\bar{\varepsilon}) = O_q(\varphi(\bar{\varepsilon})),$$

where the last step uses $\log(\mathrm{e}+2z) = O(\log(\mathrm{e}+z))$ for $z \ge 0$. Therefore, $u\bar{\varepsilon} = O_q(\varphi(\bar{\varepsilon}))$ holds as claimed. Consequently, by $\varepsilon \le \bar{\varepsilon}$ and the monotonicity of φ, we obtain

$$\varphi^\star(u) = u\bar{\varepsilon} + \varphi(\bar{\varepsilon}) = O_q(\varphi(\varepsilon)).$$

Since $\varphi^\star(\mathrm{OPT}_T+\lambda)-\varphi^\star(\lambda) \le \varphi^\star(\mathrm{OPT}_T+\lambda)$, the parameter-change contribution is $O_q(C_{\mathrm{pc}}\varphi(\varepsilon))$. Combining this with the $H_T\varphi(\varepsilon)/2$ bound above yields

$$\mathrm{Inc}_T = O_q((H_T + C_{\mathrm{pc}})\varphi(\varepsilon)).$$

Therefore, for constant $q > 0$ and $C_{\mathrm{pc}} = O(1)$, this recovers the fixed-ε inconsistency bound of $O(\varphi(\varepsilon)\log T)$ obtained by Dong and Yoshida [19]. □

G On a naive doubling reduction

We explain why we do not use a naive doubling reduction based on the fixed-ε guarantee of Dong and Yoshida [19] as a substitute for our time-varying approach in Section 4. For a fixed value of ε, their batch-to-online transformation yields a $(1+\varepsilon)$-approximate regret guarantee; equivalently, at the level of standard regret, the bound has the form of $\mathcal{R}_T \lesssim \varepsilon\mathrm{OPT}_T + H_T\varphi(\varepsilon)$ up to constants. Thus, it is natural to ask whether one can recover a small-loss regret bound by a phase-wise doubling trick that guesses the unknown scale of the phase-local offline optimum and runs the fixed-ε algorithm separately in each phase. The obstruction described below is not the deterministic algebra of such a doubling argument, but the random-order exchangeability step needed to justify applying a fixed-horizon guarantee inside phases whose endpoints are data-dependent.

Consider a threshold-based phase scheme, analogous to the standard doubling trick for an unknown scale parameter (see, e.g., Cesa-Bianchi and Lugosi [13, Exercise 2.10]). Conditional on the history before a phase starts, the unrevealed data points are still in uniformly random order. This observation alone, however, does not justify applying the fixed-horizon proof to a phase with a random endpoint, as we now explain. In a phase with threshold B, one fixes a parameter ε as if the phase-local optimum were of order B, restarts the batch-to-online procedure using only the observations in the current

phase, and stops the phase when the phase-local optimum first exceeds B. Let $T_{\mathrm{rem}} \in \mathbb{Z}_{>0}$ be the number of data points remaining when the phase starts, and write $y_1, \ldots, y_{T_{\mathrm{rem}}}$ for their random order. For $s \in \{0, \ldots, T_{\mathrm{rem}}\}$, let $Y_s = (y_1, \ldots, y_s)$ denote the phase-local prefix, with $Y_0 \coloneqq \emptyset$ and $\mathrm{OPT}(Y_0) = 0$. Define the stopping time

$$\tau \coloneqq \inf\{\, s \in \{0, \ldots, T_{\mathrm{rem}}\} \,:\, \mathrm{OPT}(Y_s) > B \,\}, \qquad \text{where} \quad \inf \emptyset = +\infty.$$

Because the losses are non-negative, $s \mapsto \mathrm{OPT}(Y_s)$ is non-decreasing; hence the phase-survival event at local time $s \in \{0, \ldots, T_{\mathrm{rem}}\}$ is equivalently

$$\{\tau > s\} = \{\mathrm{OPT}(Y_s) \le B\}.$$

Fixing ε within the phase would remove the parameter-change term inside that phase, but the proof would still have to justify the single-step random-order comparison after conditioning on the survival event $\{\tau > s\}$.

To see the issue, recall the exchangeability step used in the proof of Lemma B.1, which also appears in the fixed-ε proof of Dong and Yoshida [19, Theorem 4.1]. For a fixed, non-random local time $s \in \{1, \ldots, T_{\mathrm{rem}} - 1\}$, sample $I \sim \mathrm{Unif}(\{1, \ldots, s\})$, set $y = y_I$, and put $U = Y_s \setminus y$. If no stopping event is imposed, then conditional on U, the deleted point y and the next point y_{s+1} are exchangeable under the random-order model. Equivalently, after also conditioning on the unordered pair of held-out points, their two possible relative orders are equally likely. Since the law of $\mathcal{A}_\varepsilon(U)$ depends only on U, this exchangeability gives the identity that allows the fresh loss on y_{s+1} to be compared with the loss on a uniformly deleted point from Y_s. This is the step used before applying average sensitivity.

In a stopped phase, however, the prediction at local time $s + 1$ is made only on the event $\{\tau > s\}$. After conditioning on the same reduced prefix $U = Y_s \setminus y$, the survival event is no longer symmetric in the deleted point and the next point. Indeed, it imposes the constraint

$$\mathrm{OPT}(U \cup \{y\}) \le B$$

on the point that has already been included in the prefix, whereas the prospective next point y_{s+1} is not subject to the same constraint before it is revealed. Consequently, the stopped analogue of the deterministic-time identity may fail:

$$\mathbb{E}[\, \ell(\mathcal{A}_\varepsilon(U), y) \,|\, U, \tau > s \,] \quad \text{need not equal} \quad \mathbb{E}[\, \ell(\mathcal{A}_\varepsilon(U), y_{s+1}) \,|\, U, \tau > s \,].$$

The corresponding equality does hold without conditioning on phase survival.

A minimal example illustrates the asymmetry. After conditioning on U, suppose that the unordered pair consisting of the deleted point and the next point is $\{a, b\}$, and that we have

$$\mathrm{OPT}(U \cup \{a\}) \le B \qquad \text{and} \qquad \mathrm{OPT}(U \cup \{b\}) > B.$$

Before conditioning on survival, the two local orderings (a, b) and (b, a) are equally likely. Let

$$\mathcal{E}_{a,b} \coloneqq \{\{y, y_{s+1}\} = \{a, b\}\}.$$

After conditioning on $\{\tau > s\}$, only the ordering in which a is already in the prefix can survive; more explicitly, we have

$$\Pr[\, y = a,\ y_{s+1} = b \,|\, U, \mathcal{E}_{a,b}, \tau > s \,] = 1 \quad \text{and} \quad \Pr[\, y = b,\ y_{s+1} = a \,|\, U, \mathcal{E}_{a,b}, \tau > s \,] = 0.$$

Thus, the uniformly deleted point from a survived prefix is biased toward candidates that keep the phase below the threshold, whereas the next point may be precisely the boundary item that causes the phase to stop. Conditioning instead on $\{\tau > s + 1\}$ would constrain the next point as well, but then the boundary round crossing the threshold would be excluded from the fixed-horizon comparison and would have to be handled separately.

As we have seen above, the difficulty is not merely the possible unit loss on the threshold-crossing rounds, whose number would be logarithmic in a standard doubling scheme. The more fundamental issue is that threshold-defined phase boundaries condition every non-boundary prediction in the phase on the survival event $\{\tau > s\}$, and this conditioning is precisely what invalidates the black-box use of the fixed-horizon leave-one-out exchangeability. Therefore, the fixed-ε guarantee cannot be applied to threshold-defined phases through a black-box conditioning argument. A valid doubling proof would need an additional stopped-process analysis that reestablishes the required leave-one-out comparison, possibly with boundary terms, or it would need to use data-independent phase boundaries. We avoid this complication in the main proof: the time-varying analysis works at deterministic global times and includes the explicit parameter-change term, which is then controlled by Assumption 9.

H Basic tools for average sensitivity analysis

We present basic tools useful for analyzing average sensitivity. First, we introduce the standard coupling lemma for the total variation distance.

Lemma H.1 (Coupling lemma, cf. [38, Sections I.2 and I.5] and [1, Section 3.2])**.** *Let $(\Omega, \mathcal{F})$ be a standard Borel space. Consider probability measures P, Q on it, and let $\mathcal{C}(P, Q)$ be the set of all couplings of (P, Q). Then, for any coupling $\Gamma \in \mathcal{C}(P, Q)$, it holds that*

$$\mathrm{TV}(P, Q) \leq \Pr_{(X,Y)\sim\Gamma}[X \neq Y].$$

Moreover, we have

$$\mathrm{TV}(P, Q) = \inf_{\Gamma\in\mathcal{C}(P,Q)} \Pr_{(X,Y)\sim\Gamma}[X \neq Y],$$

and the infimum is attained by an optimal coupling.

The following lemmas are consequences of the coupling lemma.

Lemma H.2 (Product decomposition)**.** *Let E be a finite set and $\{(\Omega_f, \mathcal{F}_f)\}_{f\in E}$ standard Borel spaces. For each $f \in E$, let P_f and Q_f be probability measures on $(\Omega_f, \mathcal{F}_f)$. Then, the total variation distance between the product measures satisfies*

$$\mathrm{TV}\left(\bigotimes_{f\in E} P_f, \bigotimes_{f\in E} Q_f\right) \leq \sum_{f\in E} \mathrm{TV}(P_f, Q_f).$$

Proof. For each $f \in E$, let Γ_f be a coupling measure of (P_f, Q_f) such that

$$\Pr_{(X_f,Y_f)\sim\Gamma_f}[X_f \neq Y_f] = \mathrm{TV}(P_f, Q_f),$$

which exists by the coupling lemma (Lemma H.1). Take these couplings independently and write $X \coloneqq (X_f)_{f\in E}$ and $Y \coloneqq (Y_f)_{f\in E}$. Their joint law couples $\bigotimes_{f\in E} P_f$ and $\bigotimes_{f\in E} Q_f$, and the union bound gives

$$\Pr[X \neq Y] \leq \sum_{f\in E} \Pr[X_f \neq Y_f] = \sum_{f\in E} \mathrm{TV}(P_f, Q_f).$$

Applying Lemma H.1 to this joint law proves the claim. □

Lemma H.3 (Conditional decomposition)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space of the form $\Omega = \mathcal{Z} \times \mathcal{W}$ equipped with the product σ-algebra, where $\mathcal{Z}$ is a countable set equipped with the discrete σ-algebra and $\mathcal{W}$ is a standard Borel space. Let P and Q be probability measures on $(\Omega, \mathcal{F})$. Let P_Z and Q_Z denote the marginals of P and Q on $\mathcal{Z}$, respectively, and let $P_{W|Z=z}$ and $Q_{W|Z=z}$ denote the corresponding conditional distributions on $\mathcal{W}$ given $Z = z \in \mathcal{Z}$. At any z with $P_Z(\{z\}) = 0$, choose $P_{W|Z=z}$ to be an arbitrary probability measure on $\mathcal{W}$, and use the same convention for Q when $Q_Z(\{z\}) = 0$. Write $P_Z(z) \coloneqq P_Z(\{z\})$ and $Q_Z(z) \coloneqq Q_Z(\{z\})$. Then, it holds that*

$$\mathrm{TV}(P, Q) \leq \mathrm{TV}(P_Z, Q_Z) + \sum_{z\in\mathcal{Z}} \min\{P_Z(z), Q_Z(z)\}\, \mathrm{TV}\big(P_{W|Z=z},\, Q_{W|Z=z}\big)\,.$$

Proof. By the coupling lemma (Lemma H.1), we take $\Gamma^\star \in \mathcal{C}(P_Z, Q_Z)$ as an optimal coupling of the marginals such that

$$\Pr_{(Z,Z')\sim\Gamma^\star}[Z \neq Z'] = \mathrm{TV}(P_Z, Q_Z).$$

Also, for each $z \in \mathcal{Z}$, let $\Gamma_z \in \mathcal{C}(P_{W|Z=z}, Q_{W|Z=z})$ be an optimal coupling satisfying

$$\Pr_{(W,W')\sim\Gamma_z}[W \neq W'] = \mathrm{TV}\big(P_{W|Z=z}, Q_{W|Z=z}\big)\,.$$

Consider a coupling $\Gamma \in \mathcal{C}(P, Q)$ constructed as follows. First, sample $(Z, Z') \sim \Gamma^\star$. If $Z \neq Z'$, sample $W \sim P_{W|Z}$ and $W' \sim Q_{W|Z'}$ independently. If $Z = Z' = z$, sample $(W, W') \sim \Gamma_z$. By construction, $(Z, W) \sim P$ and $(Z', W') \sim Q$ hold. Under this coupling, we have

$$\Pr_{(Z,W,Z',W')\sim\Gamma}\big[(Z, W) \neq (Z', W')\big] \leq \Pr_{(Z,Z')\sim\Gamma^\star}\big[Z \neq Z'\big] + \Pr_{(Z,W,Z',W')\sim\Gamma}\big[Z = Z',\, W \neq W'\big].$$

The first term equals $\mathrm{TV}(P_Z, Q_Z)$ by the definition of $\Gamma^\star$. For the second term, conditioning on $Z = Z' = z$ yields

$$\Pr_{(Z,W,Z',W')\sim\Gamma}\big[Z = Z',\, W \neq W'\big] = \sum_{z\in\mathcal{Z}} \Pr_{(Z,Z')\sim\Gamma^\star}[Z = Z' = z]\,\mathrm{TV}\big(P_{W|Z=z}, Q_{W|Z=z}\big)\,.$$

For any coupling of (P_Z, Q_Z), the diagonal mass at z is at most $\min\{P_Z(z), Q_Z(z)\}$, hence

$$\Pr_{(Z,Z')\sim\Gamma^\star}[Z = Z' = z] \le \min\{P_Z(z), Q_Z(z)\}.$$

Combining the above bounds with Lemma H.1 gives

$$\begin{aligned}\mathrm{TV}(P,Q) &\le \Pr_{(Z,W,Z',W')\sim\Gamma}\big[(Z,W)\neq(Z',W')\big]\\ &\le \mathrm{TV}(P_Z,Q_Z) + \sum_{z\in\mathcal{Z}} \min\{P_Z(z),Q_Z(z)\}\,\mathrm{TV}\big(P_{W|Z=z},Q_{W|Z=z}\big)\,.\end{aligned}$$

This completes the proof. □

The following bound on the total variation distance between two uniform distributions is also convenient in our analysis.

Lemma H.4. *Let $B, B', \varepsilon > 0$, and consider uniform distributions defined on $[B,(1+\varepsilon)B]$ and $[B',(1+\varepsilon)B']$, denoted by $P = \mathrm{Unif}([B,(1+\varepsilon)B])$ and $Q = \mathrm{Unif}([B',(1+\varepsilon)B'])$, respectively. Then, we have*

$$\mathrm{TV}(P,Q) \le \frac{1+\varepsilon}{\varepsilon}\left|1 - \frac{B'}{B}\right|.$$

Proof. See the proof of Kumabe and Yoshida [35, Lemma 2.3]. □

I Randomized rounding of parameter-dependent sample counts

Some of our offline algorithms use a number of samples that depends on the approximation parameter. To establish parameter-change stability, we compare their sampling transcripts on the same input at different parameter values. Here, a sampling transcript contains the sample count and all sampling and reweighting randomness before deterministic post-processing.

Deterministic rounding of the sample count creates a difficulty: arbitrarily close parameter values can yield different counts, making the transcript distributions disjoint and their total variation distance equal to one. We therefore use randomized rounding, which allows us to bound the change in the count distribution by the change in its mean.

For a target mean sample count $\mu \ge 0$, draw $U \sim \mathrm{Unif}([0,1))$ and define

$$R_\mu := \lfloor \mu + U \rfloor.$$

Write $N \sim R_\mu$ when N has the same distribution as R_μ. We use this count before taking independent samples. For the running-time guarantees below, each draw from an explicitly specified finite or uniform distribution is treated as a unit-cost operation. For a random variable Z, write $\mathcal{L}(Z)$ for its probability law.

Lemma I.1 (Properties of randomly rounded sample counts)**.** *For every $\mu, \nu \ge 0$, we have*

$$R_\mu \in \{\lfloor\mu\rfloor, \lceil\mu\rceil\}, \qquad \mathbb{E}[R_\mu] = \mu, \qquad R_\mu \le \lceil\mu\rceil \quad \textit{almost surely},$$

and

$$\mathrm{TV}(\mathcal{L}(R_\mu), \mathcal{L}(R_\nu)) \le \min\{1, |\mu-\nu|\} \le |\mu-\nu|.$$

Proof. Write $\mu = m + r$, where $m := \lfloor\mu\rfloor$ and $r \in [0,1)$. Then we have $R_\mu = m$ with probability $1-r$ and $R_\mu = m+1$ with probability r. This proves the support and almost-sure bounds, and we have $\mathbb{E}[R_\mu] = m(1-r) + (m+1)r = \mu$.

For the total variation bound, couple R_μ and R_ν using the same uniform variable U. If $|\mu-\nu| \ge 1$, their mismatch probability is at most one. If $|\mu-\nu| < 1$, the set

$$\{\, U \in [0,1) \,:\, \lfloor \mu + U \rfloor \neq \lfloor \nu + U \rfloor \,\}$$

is an interval modulo one of Lebesgue measure $|\mu-\nu|$. The coupling inequality in Lemma H.1 proves the claim. □

Lemma I.2 (Total variation bound for sampling transcripts)**.** *For each $m \in \mathbb{Z}_{\geq 0}$, let $\mathcal{T}_m$ be a count-m transcript space, and let P_m and Q_m be probability laws on $\mathcal{T}_m$. Let $\mathsf{Mix}(\mu, \{P_m\}_{m\geq 0})$ denote the law on the disjoint union of the count-indexed transcript spaces obtained by first drawing $N \sim R_\mu$ and then drawing a transcript from P_N. Then, for any $\mu, \nu \geq 0$, we have*

$$\mathrm{TV}(\mathsf{Mix}(\mu, \{P_m\}_{m\geq 0}), \mathsf{Mix}(\nu, \{Q_m\}_{m\geq 0})) \leq |\mu - \nu| + \mathbb{E}_{N\sim R_\mu}[\mathrm{TV}(P_N, Q_N)].$$

Proof. Apply Lemma H.3 to the count coordinate and use Lemma I.1 for the count marginals. Writing $a_m := \Pr[R_\mu = m]$ and $b_m := \Pr[R_\nu = m]$, the remaining conditional term is bounded by

$$\sum_{m\geq 0} \min\{a_m, b_m\}\, \mathrm{TV}(P_m, Q_m) \leq \mathbb{E}_{N\sim R_\mu}[\mathrm{TV}(P_N, Q_N)].$$

Thus, we have the desired bound. □

Lemma I.3 (Preserving failure-probability bounds under randomized rounding)**.** *Fix $\varepsilon \in (0, 1]$ and $\delta \in [0, 1]$, a mean sample count $\mu(\varepsilon) \in \mathbb{R}_{\geq 0}$, and a threshold $m_0(\varepsilon) \in \mathbb{R}_{\geq 0}$. Suppose that a fixed-size sampling argument has failure probability at most δ uniformly for every integer sample size $s \geq m_0(\varepsilon)$. If $N_\varepsilon \sim R_{\mu(\varepsilon)}$ and $\lfloor \mu(\varepsilon) \rfloor \geq m_0(\varepsilon)$, then the randomly rounded routine has failure probability at most δ.*

Proof. Every realization satisfies $N_\varepsilon \geq \lfloor \mu(\varepsilon) \rfloor \geq m_0(\varepsilon)$. Conditioning on the realized count and applying the corresponding fixed-size guarantee proves the claim. □

Lemma I.4 (Uniform upper bound on sample count)**.** *Let $\varepsilon_t := \varepsilon_{\min}(\mathrm{OPT}_t + \lambda)$, where the losses lie in $[0, 1]$ and $0 < \lambda \leq 1$, and let $\mu\colon (0, 1] \to \mathbb{R}_{\geq 0}$ be nonincreasing. For every online history, we have*

$$N_t \sim R_{\mu(\varepsilon_t)} \quad \Longrightarrow \quad N_t \leq \lceil \mu(\varepsilon_{\min}(T + \lambda)) \rceil \quad \textit{almost surely}.$$

Proof. Because $0 \leq \mathrm{OPT}_t \leq t \leq T$ and $\varepsilon_{\min}$ is nonincreasing, we have $\varepsilon_t \geq \varepsilon_{\min}(T + \lambda)$. The monotonicity of μ and the support bound $R_\nu \leq \lceil \nu \rceil$ in Lemma I.1 give the claim. □

We will also use the following bound on the total variation distance between two uniform distributions with different parameters.

Lemma I.5 (Uniform perturbations with different parameters)**.** *Let $B, B' > 0$ and $0 < \varepsilon \leq \eta \leq 1$. Let $P = \mathrm{Unif}([B, (1 + \varepsilon)B])$ and $Q = \mathrm{Unif}([B', (1 + \eta)B'])$. Then, it holds that*

$$\mathrm{TV}(P, Q) = O\left(\frac{1}{\varepsilon}\left|1 - \frac{B'}{B}\right| + \frac{\eta - \varepsilon}{\varepsilon}\right).$$

Proof. Consider using the triangle inequality to compare P and Q through the intermediate distribution $Q' = \mathrm{Unif}([B', (1 + \varepsilon)B'])$. The first comparison of P and Q' is controlled by Lemma H.4. For the second comparison of Q' and Q, the left endpoint is fixed and only the interval length changes; the overlap ratio is ε/η, and thus the total variation distance is $1 - \varepsilon/\eta \leq (\eta - \varepsilon)/\varepsilon$. □

Sampling with randomized weights. We give a common stability analysis for the weighted sampling procedures used in Appendices L, M.1 and M.2. These procedures randomize inverse-probability weights in one of two ways, by perturbing the denominator or by drawing the weight directly. Let r be a probability vector on a finite label set, let $\mu \geq 0$, and let $0 < \beta \leq 1$. First draw $N \sim R_\mu$. Conditional on $N = s > 0$, draw independent labels $I_1, \ldots, I_s$ from r. For each sampled label $I_j = i$ with $r_i > 0$, consider two ways to randomize the weight $1/(sr_i)$, using independent randomness across $j \in \{1, \ldots, s\}$ conditional on the labels:

(i) Draw V_j uniformly from $[r_i, (1 + \beta)r_i]$ and assign weight $1/(sV_j)$ to the sampled label.

(ii) Draw W_j uniformly from $[1/(sr_i), (1 + \beta)/(sr_i)]$ and assign weight W_j to the sampled label.

Let $\mathcal{T}^{\mathrm{den}}_{\mu,\beta}(r)$ denote the law of $(N, ((I_j, V_j))_{j=1}^N)$ under denominator perturbation in (i), and let $\mathcal{T}^{\mathrm{wt}}_{\mu,\beta}(r)$ denote the law of $(N, ((I_j, W_j))_{j=1}^N)$ under direct weight perturbation in (ii). Each transcript contains the count and an ordered sequence of that many label–real-number pairs. At $s = 0$, use the empty sequence; for $r_i = 0$, take the conditional law of V_j or W_j given $I_j = i$ to be the point mass at zero.

To obtain a transcript that directly includes the weights assigned in (i), retain the count N and replace each pair (I_j, V_j) by $(I_j, 1/(NV_j))$, using weight zero if $NV_j = 0$ and the empty sequence if $N = 0$. This map is measurable and independent of r, μ, β; by data processing, it cannot increase total variation distances between laws of the transcripts in (i). The transcript in (ii) already contains the assigned weights.

Lemma I.6. *Let p, q be probability vectors on the same finite label set, let $\mu \geq 0$, and let $0 < \beta \leq 1$. For the two transcript laws defined above, we have*

$$\mathrm{TV}\big(\mathcal{T}^{\mathrm{den}}_{\mu,\beta}(p), \mathcal{T}^{\mathrm{den}}_{\mu,\beta}(q)\big) \leq \frac{2+\beta}{2\beta}\mu\|p-q\|_1 \qquad \textit{and} \qquad \mathrm{TV}\big(\mathcal{T}^{\mathrm{wt}}_{\mu,\beta}(p), \mathcal{T}^{\mathrm{wt}}_{\mu,\beta}(q)\big) \leq \frac{3\mu}{\beta}\|p-q\|_1.$$

Proof. Let $\mathcal{I}$ be the common finite label set. For either procedure, fix a count $s \in \mathbb{Z}_{>0}$. By independence, Lemma H.2 bounds the total variation distance between the count-s transcript laws under p and q by s times the one-draw distance. When $N = 0$, the transcript under either p or q consists of the count zero and an empty sequence. By Lemma I.2 and $\mathbb{E}[N] = \mu$, it therefore suffices to prove the two claimed bounds with μ replaced by 1, uniformly in s.

For an interval $J \subseteq \mathbb{R}$ and $v \in \mathbb{R}$, the indicator $\mathbf{1}_J(v)$ equals 1 when $v \in J$ and 0 otherwise. For the first procedure, let P and Q denote the joint laws of (I_1, V_1) under label probabilities p and q, respectively. For each $r \in \{p, q\}$, the conditional interval at a label $i \in \mathcal{I}$ with $r_i > 0$ has length βr_i. Multiplying its uniform density by the label probability gives $r_i/(\beta r_i) = 1/\beta$. With respect to counting measure on $\mathcal{I}$ and Lebesgue measure on $\mathbb{R}$, we therefore use the joint density f_r, defined for $i \in \mathcal{I}$ and $v \in \mathbb{R}$ by

$$f_r(i, v) := \begin{cases} \beta^{-1}\mathbf{1}_{[r_i,(1+\beta)r_i]}(v), & r_i > 0, \\ 0, & r_i = 0. \end{cases}$$

Using the joint densities f_p and f_q, we have

$$\begin{aligned} \mathrm{TV}(P, Q) &= \frac{1}{2}\sum_{i\in\mathcal{I}}\int_{\mathbb{R}} |f_p(i, v) - f_q(i, v)|\, dv \\ &\leq \frac{1}{2\beta}\sum_{i\in\mathcal{I}}\big(|p_i - q_i| + |(1+\beta)p_i - (1+\beta)q_i|\big) \\ &= \frac{2+\beta}{2\beta}\|p-q\|_1. \end{aligned}$$

For the second procedure, let P and Q denote the joint laws of (I_1, W_1) under label probabilities p and q, respectively. For each $i \in \mathcal{I}$ with $p_i q_i > 0$, Lemma H.4 bounds the distance between the two conditional laws of W_1 given $I_1 = i$ by $(1+\beta)|p_i - q_i|/(\beta \max\{p_i, q_i\})$. Coordinates with zero common probability make no conditional contribution. Applying Lemma H.3, we have

$$\begin{aligned} \mathrm{TV}(P, Q) &\leq \frac{1}{2}\|p-q\|_1 + \sum_{\substack{i\in\mathcal{I} \\ p_i q_i > 0}} \min\{p_i, q_i\}\frac{(1+\beta)|p_i - q_i|}{\beta \max\{p_i, q_i\}} \\ &\leq \left(\frac{1}{2} + \frac{1+\beta}{\beta}\right)\|p-q\|_1 \leq \frac{3}{\beta}\|p-q\|_1. \end{aligned}$$

The second inequality uses $\min\{p_i, q_i\}/\max\{p_i, q_i\} \leq 1$ on the summation domain, and the last uses $\beta \leq 1$. The above inequalities give the required one-draw bounds uniformly in s, completing the proof. □

We next keep the sampling probabilities fixed and allow the mean sample count and the perturbation parameter to change with the target accuracy.

Lemma I.7. *Fix a probability vector p independent of the accuracy parameter and $c \in (0,1]$. Write $\mathcal{T}$ for either $\mathcal{T}^{\mathrm{den}}$ or $\mathcal{T}^{\mathrm{wt}}$, using the same procedure at both accuracy parameters. For $0 < \varepsilon \le \eta \le 1$ and arbitrary nonnegative mean counts $\mu_\varepsilon, \mu_\eta$, set the perturbation parameter β to $c\varepsilon$ at accuracy ε and to $c\eta$ at accuracy η. Then we have*

$$\mathrm{TV}\big(\mathcal{T}_{\mu_\varepsilon, c\varepsilon}(p), \mathcal{T}_{\mu_\eta, c\eta}(p)\big) \le |\mu_\varepsilon - \mu_\eta| + \mu_\varepsilon \frac{\eta - \varepsilon}{\varepsilon}. \tag{5}$$

In particular, if $\mu_\alpha = C\alpha^{-a}\log(\mathrm{e} + B/\alpha) + b$ for $0 < \alpha \le 1$, with fixed $C \ge 1$, $a > 0$, $B \ge 0$, and additive constant $b \ge 0$ independent of α, then the right-hand side of (5) is at most $(a+2)\mu_\varepsilon(\eta-\varepsilon)/\varepsilon$.

Proof. For either procedure, fix a count $s \in \mathbb{Z}_{\ge 0}$. For $s > 0$ and each label i with $p_i > 0$, the conditional laws given $I_1 = i$ at accuracy $\alpha \in \{\varepsilon, \eta\}$ are

$$\mathrm{Unif}([p_i, (1 + c\alpha)p_i]) \quad \text{in (i)}, \qquad \mathrm{Unif}\left(\left[\frac{1}{sp_i}, \frac{1 + c\alpha}{sp_i}\right]\right) \quad \text{in (ii)}.$$

In either procedure, the interval at ε is contained in the interval at η, with the same left endpoint and length ratio ε/η. The total variation distance between the two conditional laws is $1 - \varepsilon/\eta$. At a label of probability zero, both conditional laws are point masses at zero. For $s = 0$, both transcripts consist of the count zero and an empty sequence. Since $R_s = s$ almost surely and the label law p is unchanged, Lemmas H.2 and H.3 give

$$\mathrm{TV}(\mathcal{T}_{s,c\varepsilon}(p), \mathcal{T}_{s,c\eta}(p)) \le s\left(1 - \frac{\varepsilon}{\eta}\right) \le s\frac{\eta - \varepsilon}{\varepsilon}.$$

Applying Lemma I.2 with this inequality and using $\mathbb{E}[R_{\mu_\varepsilon}] = \mu_\varepsilon$ proves (5).

For the stated choice of μ_α, differentiation gives

$$0 \le -\frac{d}{d\alpha}\mu_\alpha \le \frac{a+1}{\alpha}\mu_\alpha \qquad (0 < \alpha \le 1).$$

Since μ_α is nonincreasing, integrating over $[\varepsilon, \eta]$ gives

$$|\mu_\varepsilon - \mu_\eta| = \int_\varepsilon^\eta \left(-\frac{d}{d\alpha}\mu_\alpha\right) d\alpha \le (a+1)\mu_\varepsilon \frac{\eta - \varepsilon}{\varepsilon}.$$

Substituting this inequality into (5) proves the last claim. □

J Regularity condition for establishing Assumption 9

We present an elementary consequence of the accuracy parameter choice given in (1) that will play a key role in verifying Assumption 9 in each application. For a sensitivity function φ differentiable on $(0,1]$, put

$$\psi(\varepsilon) := -\varphi'(\varepsilon) \qquad (0 < \varepsilon \le 1),$$

where the value at $\varepsilon = 1$ is understood as the left-derivative value. We use the following three primitive conditions: there exist constants $A, \rho, B > 0$ such that

$$\varphi(\varepsilon) \le A\varepsilon\psi(\varepsilon) \quad \text{for all } 0 < \varepsilon \le 1, \tag{6}$$

$$\psi(\eta) \le \psi(\varepsilon)\left(\frac{\varepsilon}{\eta}\right)^\rho \quad \text{for all } 0 < \varepsilon \le \eta \le 1, \tag{7}$$

$$\psi(1) \ge B. \tag{8}$$

The first condition says that the value of φ is controlled by its marginal cost $\psi(\varepsilon) = -\varphi'(\varepsilon)$ multiplied by ε, while the second condition says that this marginal cost is power-regular under multiplicative changes of the accuracy parameter. The third one is a technical endpoint condition. The following lemma offers a convenient calculus for verifying Assumption 9.

Lemma J.1. *Let $\varphi\colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0} \cup \{+\infty\}$ satisfy $\varphi(\varepsilon) = +\infty$ outside $(0,1]$, be finite and continuously differentiable on $(0,1]$, and satisfy $\lim_{\varepsilon \downarrow 0} \varphi(\varepsilon) = +\infty$. Assume that $\psi(\varepsilon) := -\varphi'(\varepsilon) > 0$*

holds on $(0, 1]$ *and that* (6)–(8) *hold for some constants* $A, \rho, B > 0$. *Let* $\varepsilon_{\min}$ *denote the parameter selection rule in* (1). *If* $u \geq v > 0$ *and* $u - v \leq 1$, *then we have* $\varepsilon_{\min}(u) \leq \varepsilon_{\min}(v)$ *and*

$$\varphi(\varepsilon_{\min}(u))\frac{\varepsilon_{\min}(v) - \varepsilon_{\min}(u)}{\varepsilon_{\min}(u)} = O_{A,\rho,B}(\varepsilon_{\min}(u)(u - v)),$$

equivalently, $\varphi(\varepsilon_{\min}(u))(\varepsilon_{\min}(v) - \varepsilon_{\min}(u)) = O_{A,\rho,B}\big(\varepsilon_{\min}(u)^2(u - v)\big)$. *In particular, if the function* φ *takes the form considered in Corollary 12, i.e.,*

$$\varphi(\varepsilon) = C_1\varepsilon^{-q}\log(\mathrm{e} + C_2/\varepsilon) \qquad (0 < \varepsilon \leq 1),$$

with $\varphi(\varepsilon) = +\infty$ *outside* $(0, 1]$, *where* $q > 0$, $C_1 \geq 1$, *and* $C_2 \geq 0$, *then* (6)–(8) *hold with* $A = 1/q$, $\rho = q + 1$, *and* $B = q$. *Consequently, for this specific form, it holds that*

$$\varphi(\varepsilon_{\min}(u))\frac{\varepsilon_{\min}(v) - \varepsilon_{\min}(u)}{\varepsilon_{\min}(u)} = O_q(\varepsilon_{\min}(u)(u - v)),$$

equivalently, $\varphi(\varepsilon_{\min}(u))(\varepsilon_{\min}(v) - \varepsilon_{\min}(u)) = O_q\big(\varepsilon_{\min}(u)^2(u - v)\big)$.

Proof. The monotonicity $\varepsilon_{\min}(u) \leq \varepsilon_{\min}(v)$ follows from the properties of $\varepsilon_{\min}$ in Lemma 7. It remains to prove the quantitative bound. We first prove the general bound under (6)–(8), and then verify these conditions for the specific form. Let

$$a \coloneqq \varepsilon_{\min}(u), \qquad b \coloneqq \varepsilon_{\min}(v), \qquad \Delta \coloneqq u - v, \qquad \text{and} \qquad r \coloneqq \frac{b - a}{a}.$$

By the monotonicity of $\varepsilon_{\min}(\cdot)$, we have $0 < a \leq b \leq 1$. If $a = 1$, then $b = 1$ holds and the claim is trivial. Hence, assume $a < 1$. For a scalar $w > 0$, write $g_w(\varepsilon) \coloneqq w\varepsilon + \varphi(\varepsilon)$ for the objective minimized in the definition of $\varepsilon_{\min}(w)$. Since $a = \varepsilon_{\min}(u)$ is an interior minimizer of g_u and $g_u'(\varepsilon) = u - \psi(\varepsilon)$, we have $u = \psi(a)$. If $b < 1$, the same argument gives $v = \psi(b)$. If $b = 1$, the one-sided optimality condition for minimizing g_v over $(0, 1]$ gives $g_v'(1) \leq 0$, equivalently $v \leq \psi(1) = \psi(b)$. Thus, in both cases, it holds that

$$\Delta \geq \psi(a) - \psi(b) = u - \psi(b). \tag{9}$$

Set $x \coloneqq b/a = 1 + r$ and $R \coloneqq (1 + B^{-1})^{1/\rho}$. By (7), we have $\psi(b) \leq ux^{-\rho}$, so (9) gives

$$\Delta \geq u(1 - x^{-\rho}).$$

By (8) and (7), together with $b \leq 1$, we have

$$B \leq \psi(1) \leq ua^{\rho} \leq u(a/b)^{\rho} = ux^{-\rho}.$$

Using $x \geq 1$ and $\Delta \leq 1$, we obtain

$$B(x^{\rho} - 1) \leq ux^{-\rho}(x^{\rho} - 1) = u(1 - x^{-\rho}) \leq \Delta \leq 1.$$

Consequently, we have $1 \leq x \leq R$. Integrating the positive decreasing function $t \mapsto \rho t^{-\rho-1}$ on $[1, x]$ gives

$$\Delta \geq u(1 - x^{-\rho}) = u\int_1^x \rho t^{-\rho-1}\,dt \geq \rho u R^{-\rho-1}(x - 1).$$

Using (6), we obtain

$$\varphi(a)r \leq Aau(x - 1) \leq \frac{AR^{\rho+1}}{\rho}a\Delta.$$

The constant depends only on A, ρ, B; multiplication by a gives the equivalent product bound. The same calculation covers the case $\Delta = 0$.

It remains to verify the three primitive conditions for $\varphi(\varepsilon) = C_1\varepsilon^{-q}\log(\mathrm{e} + C_2/\varepsilon)$. Set

$$L(\varepsilon) = \log(\mathrm{e} + C_2/\varepsilon), \qquad R(\varepsilon) = \frac{C_2}{\mathrm{e}\varepsilon + C_2}, \qquad \text{and} \qquad M(\varepsilon) = qL(\varepsilon) + R(\varepsilon).$$

Then, we have

$$\varphi(\varepsilon) = C_1\varepsilon^{-q}L(\varepsilon) \qquad \text{and} \qquad \psi(\varepsilon) = -\varphi'(\varepsilon) = C_1\varepsilon^{-q-1}M(\varepsilon).$$

Since $R(\varepsilon) \geq 0$, we have $M(\varepsilon) \geq qL(\varepsilon)$. Thus, from $L(\varepsilon) \geq 1$, it follows that

$$\frac{\varphi(\varepsilon)}{\varepsilon\psi(\varepsilon)} = \frac{L(\varepsilon)}{M(\varepsilon)} \leq \frac{1}{q}.$$

Hence (6) holds with $A = 1/q$. Moreover, both L and R are non-increasing on $(0, 1]$, and thus M is non-increasing. Therefore, if $0 < \varepsilon \leq \eta \leq 1$, then we have

$$\psi(\eta) = C_1\eta^{-q-1}M(\eta) \leq C_1\eta^{-q-1}M(\varepsilon) = \psi(\varepsilon)\left(\frac{\varepsilon}{\eta}\right)^{q+1}.$$

Thus, (7) holds with $\rho = q + 1$. Finally, it holds that

$$\psi(1) = C_1(qL(1) + R(1)) \geq q,$$

because $C_1 \geq 1$, $L(1) \geq 1$, and $R(1) \geq 0$. Thus (8) holds with $B = q$. Plugging $A = 1/q$, $\rho = q + 1$, and $B = q$ into the general bound completes the proof. □

A convenient sufficient template for checking Lemma J.1 is the following. Suppose

$$\varphi(\varepsilon) = C\varepsilon^{-q}L(\varepsilon), \qquad C \geq 1, \qquad \text{and} \qquad q > 0,$$

where $L \geq 1$ is continuously differentiable and non-increasing on $(0, 1]$. If

$$M(\varepsilon) \coloneqq qL(\varepsilon) - \varepsilon L'(\varepsilon)$$

is non-increasing, then $\psi(\varepsilon) = C\varepsilon^{-q-1}M(\varepsilon)$ satisfies (6) with $A = 1/q$ and (7) with $\rho = q + 1$. Moreover, since $M(1) \geq qL(1) \geq q$, condition (8) holds with $B = q$. Thus the proof above applies without using any special property of the logarithm beyond these monotonicity checks.

K Offline algorithm for submodular function minimization

We detail the offline algorithm for submodular function minimization and prove the average sensitivity bound stated in Proposition 13. Below, we work under value-oracle access to each splitting submodular function as in Kenneth and Krauthgamer [34], while we assume the exact-real model to handle perturbation by uniform noise introduced later. Throughout the anchored-lift discussion, V denotes the original ground set and $n = |V|$. In the generic sparsification analysis, we write the vertex set as U. In the anchored setting, $U = V^+$ and hence $|U| = n + 1$, which does not change the asymptotic bounds after translating back to the original ground-set size.

K.1 Anchored lift for non-normalized submodular losses

The cut-sparsifier theorem of Kenneth and Krauthgamer [34] is stated for splitting functions normalized at the empty set. The online submodular losses in Section 5 need not satisfy $\ell_t(\emptyset) = 0$. We therefore use the following anchored representation.

Let V be the original ground set and let $r \notin V$ be a dummy vertex. For this anchored lift, set $U \coloneqq V^+ = V \cup \{r\}$. For a non-negative submodular function $\ell\colon 2^V \to \mathbb{R}_{\geq 0}$, define $g_\ell\colon 2^U \to \mathbb{R}_{\geq 0}$ by

$$g_\ell(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ \ell(A \cap V) & \text{if } A \neq \emptyset. \end{cases}$$

A value query to g_ℓ is implemented by either returning zero, when the queried set is empty, or by making one value query to the original oracle for ℓ on $A \cap V$. Thus the lift is compatible with the same value-oracle model as the original losses.

Lemma K.1. *The function g_ℓ is non-negative and submodular on U, and it satisfies $g_\ell(\emptyset) = 0$. Moreover, for every $S \subseteq V$, we have $g_\ell(S \cup \{r\}) = \ell(S)$.*

Proof. Non-negativity and the identity on anchored sets are immediate from the definition. It remains to check submodularity. Take arbitrary $A, B \subseteq U$. If $A = \emptyset$ or $B = \emptyset$, the submodularity inequality for g_ℓ holds with equality. Assume next that $A, B \neq \emptyset$ and $A \cap B \neq \emptyset$. Then, we have

$$g_\ell(A)+g_\ell(B) = \ell(A\cap V)+\ell(B\cap V) \geq \ell((A\cup B)\cap V)+\ell((A\cap B)\cap V) = g_\ell(A\cup B)+g_\ell(A\cap B),$$

where the inequality is the submodularity of ℓ. Finally, suppose that $A, B \neq \emptyset$ and $A \cap B = \emptyset$. Since $g_\ell(A \cap B) = 0$ and $\ell(\emptyset) \geq 0$, submodularity of ℓ gives

$$g_\ell(A) + g_\ell(B) = \ell(A \cap V) + \ell(B \cap V) \geq \ell((A \cup B) \cap V) + \ell(\emptyset) \geq g_\ell(A \cup B) + g_\ell(A \cap B).$$

Thus, g_ℓ is submodular. □

For a multiset X of submodular loss functions, let $H_X^+ = (U, E_X, \{g_e\}_{e \in E_X})$ be the lifted hypergraph obtained by creating one hyperedge $e = U$ for each $\ell \in X$ and setting $g_e = g_\ell$. The original objective is exactly represented on anchored cuts:

$$\sum_{\ell \in X} \ell(S) = \mathrm{cut}_{H_X^+}(S \cup \{r\}) \qquad \text{for every } S \subseteq V.$$

After constructing a sparsifier $\tilde{H}_X^+$, we do not minimize over all cuts of $U = V^+$. Instead, we define an ordinary set function on the original ground set by

$$F_{\tilde{H}_X^+}(S) \coloneqq \mathrm{cut}_{\tilde{H}_X^+}(S \cup \{r\}) = \sum_{e \in \tilde{E}_X} w_e \, g_e((S \cup \{r\}) \cap e) \qquad \text{for every } \; S \subseteq V,$$

where $\tilde{E}_X$ is the sampled hyperedge multiset and $w_e \geq 0$ denotes the sampled reweighting factor. For each sampled hyperedge e, let ℓ_e be its original loss. Each sampled hyperedge is U, so we have $g_e((S \cup \{r\}) \cap e) = \ell_e(S)$. Thus $F_{\tilde{H}_X^+}$ is a non-negative weighted sum of the original non-negative submodular losses, so it is submodular on V. A value query is answered by querying each sampled loss oracle at S and taking the weighted sum. The post-processing step is therefore value-oracle submodular function minimization on V, with a fixed deterministic tie-breaking rule. Note that, although the ambient lifted hypergraph has the empty cut $A = \emptyset$, the post-processing step minimizes only over anchored cuts $A = S \cup \{r\}$ over $S \subseteq V$. Thus, $A = \emptyset$ is not feasible, and the original empty action corresponds to $\{r\}$, not to $\emptyset$.

From the next subsection onward, $H = (U, E, \{g_e\}_{e \in E})$ denotes a generic submodular hypergraph. When these statements are applied to the anchored construction above, we take $U = V^+$ and restrict the final minimization to anchored cuts $S \cup \{r\}$ with $S \subseteq V$.

K.2 Sparsification via importance sampling

We describe the importance-sampling-based algorithm of Kenneth and Krauthgamer [34] in detail. Recall that a submodular hypergraph is given by $H = (U, E, \{g_e\}_{e \in E})$ with vertex set U, hyperedges E, and non-negative submodular splitting functions $g_e \colon 2^e \to \mathbb{R}_{\geq 0}$ satisfying $g_e(\emptyset) = 0$ for each $e \in E$. Define

$$g^e_{a \to b} \coloneqq \min_{A \subseteq U \colon a \in A, \; b \notin A} g_e(A \cap e) \qquad \text{for distinct } a, b \in U.$$

In what follows, all sums over ordered pairs are taken over distinct pairs. Following Kenneth and Krauthgamer [34], define

$$\rho_e(H) \coloneqq \sum_{\substack{(a,b) \in U \times U \\ a \neq b}} \frac{g^e_{a \to b}}{\sum_{f \in E} g^f_{a \to b}}, \quad \rho'_e(H) \coloneqq \frac{g_e(e)}{\sum_{f \in E} g_f(f)} \quad \text{and} \quad s_e(H) \coloneqq \rho_e(H) + \rho'_e(H),$$

with the convention that zero denominators yield zero ratios. Note that we have

$$\sum_{e \in E} \rho_e(H) \leq |U|^2 \qquad \text{and} \qquad \sum_{e \in E} \rho'_e(H) \leq 1. \tag{10}$$

Let $M > 0$, which is specified later. The sparsification algorithm of Kenneth and Krauthgamer [34] samples each $e \in E$ independently with probability

$$p_e(H) = \min\{1, \, M s_e(H)\},$$

sets $g'_e \equiv g_e / p_e(H)$ for each selected hyperedge $e \in E$, and returns the resulting sparsifier.

The fixed-cut concentration bound in the proof of Kenneth and Krauthgamer [34, Theorem 1.4] is $2\exp(-c\varepsilon^2 M)$ for an absolute constant $c > 0$. A union bound over the at most $2^{|U|}$ cuts shows that, for every $\delta \in (0, 1)$, taking $M = \Theta(\varepsilon^{-2}(|U| + \log(1/\delta)))$ gives a $(1 \pm \varepsilon)$-sparsifier with probability at least $1 - \delta$, with $\mathbb{E}[|E'|] \leq M(|U|^2 + 1)$. We use $\delta = \delta_T = (2T)^{-1}$; since a size-$t$ prefix has total loss at most t, the failure event contributes at most $t\delta_T = O(t/T)$.

Remark K.2 (Accuracy translation)**.** Let $0 \le \xi < 1$ and $\beta \ge 0$, and let F, $\bar{F}$, and $\widetilde{F}$ be nonnegative objectives on the same feasible set. Suppose that $\bar{F}$ is a two-sided $(1 \pm \xi)$ approximation to an original objective F, and that the objective $\widetilde{F}$ used for optimization satisfies $\bar{F}/(1+\beta) \le \widetilde{F} \le \bar{F}$ pointwise. If $\widehat{\theta}$ minimizes $\widetilde{F}$ and $\theta^\star$ minimizes F, then we have

$$F(\widehat{\theta}) \le \frac{(1+\beta)(1+\xi)}{1-\xi} F(\theta^\star).$$

For target accuracy $\varepsilon \in (0, 1]$, the choices $\xi = \varepsilon/12$ and $\beta = \varepsilon/2$ satisfy

$$\frac{(1+\varepsilon/2)(1+\varepsilon/12)}{1-\varepsilon/12} \le 1+\varepsilon.$$

Computational complexity. We work in the exact-real value-oracle model stated in the preliminaries. We compute every directed minimum used to form the importance scores and the final minimizer, with lexicographic tie-breaking under a fixed ordering of V, using polynomially many calls to a strongly polynomial submodular-function-minimization algorithm [29, 32, 48]. All other score-computation, sampling, reweighting, and continuous-perturbation steps use polynomially many unit-cost real arithmetic operations, comparisons, random draws, and loss-oracle queries. The expected number of hyperedges in the resulting sparsifier is $\tilde{O}(\varepsilon^{-2}|U|^3)$. In the anchored reduction, we have $|U| = n+1$, and Appendix K.1 shows that the optimized function is an ordinary submodular function on the original ground set and that its value oracle is implemented by summing value-oracle queries to the sampled losses. Hence, for every fixed $\varepsilon \in (0, 1]$, the offline $(1+\varepsilon)$-approximation algorithm runs in polynomial time in the stated exact-real value-oracle model. The complexity of the adaptive online procedure is bounded as follows. At prefix t, both the prefix and sampled objectives are ordinary submodular functions whose value oracles use at most t loss-oracle calls. Moreover, each input hyperedge is sampled at most once, and thus the sampled edge multiset E'_t satisfies $E'_t \subseteq E_t$ and $|E'_t| \le t \le T$, independently of ε_t; therefore, no lower bound on ε_t is needed. Combining these facts with the polynomial-time construction above shows that every round runs in polynomial time in the exact-real value-oracle model.

For later use, we present basic lemmas on this importance-sampling-based sparsification.

Lemma K.3. *For $x, y \ge 0$ and $M > 0$, $|\min\{1, Mx\} - \min\{1, My\}| \le M|x-y|$ holds.*

Proof. The map $x \mapsto \min\{1, Mx\}$ is piecewise linear with slope in $[0, M]$, hence M-Lipschitz. □

Lemma K.4. *For any $e \in E$, we have $\sum_{f \in E,\ f \ne e} |s_f(H) - s_f(H-e)| \le s_e(H)$.*

Proof. First, we have

$$\sum_{f \in E,\ f \ne e} |s_f(H) - s_f(H-e)| \le \sum_{f \in E,\ f \ne e} |\rho_f(H) - \rho_f(H-e)| + \sum_{f \in E,\ f \ne e} |\rho'_f(H) - \rho'_f(H-e)|.$$

We bound the two parts separately. Fix distinct $a, b \in U$ and $e \in E$ and abbreviate $D = \sum_{f \in E} g^f_{a \to b}$ and $\alpha = g^e_{a \to b}$. If $D = 0$ or $D = \alpha$, we have $g^f_{a \to b} = 0$ for every $f \in E$ with $f \ne e$, and the contribution of this ordered pair to $\rho_f(H) - \rho_f(H-e)$ is zero by the zero-denominator convention. Hence, we may assume that $D > 0$ and $\alpha < D$. It holds that

$$\sum_{f \in E,\ f \ne e} \left| \frac{g^f_{a \to b}}{D - \alpha} - \frac{g^f_{a \to b}}{D} \right| = \sum_{f \in E,\ f \ne e} \left(\frac{g^f_{a \to b}}{D - \alpha} - \frac{g^f_{a \to b}}{D} \right) = 1 - \frac{D - \alpha}{D} = \frac{\alpha}{D}.$$

Summing over distinct ordered pairs (a, b) and using $|\sum_i x_i| \le \sum_i |x_i|$ yield the claim for the ρ part. The same argument with $D = \sum_{f \in E} g_f(f)$ and $\alpha = g_e(e)$ bounds the ρ' part. Consequently, the desired inequality follows by adding the two parts. □

K.3 Weight perturbation

We introduce a small random perturbation to the weights of sampled hyperedges, thereby controlling the average sensitivity of the full output, including both sampled hyperedges and assigned weights.

This idea is already studied in Yoshida and Ito [63] and Dong and Yoshida [19], but we restate it here for completeness.

Recall that we sample each hyperedge $e \in E$ independently with probability

$$p_e(H) = \min\{1,\, Ms_e(H)\},$$

and, if selected, rescale g_e by the weight $1/p_e(H)$.

For target optimization accuracy $\varepsilon \in (0, 1]$, set $\xi := \varepsilon/12$ and $\beta := \varepsilon/2$, as in Remark K.2. By setting $M = \Theta(\xi^{-2}(|U| + \log T)) = \Theta(\varepsilon^{-2}(|U| + \log T))$, with probability at least $1 - \delta_T$, the original construction yields a $(1 \pm \xi)$-cut sparsifier of H.

Perturbation rule. Whenever a hyperedge e is sampled, instead of using the deterministic weight $1/p_e(H)$, we draw an independent random variable

$$\widetilde{p}_e \sim \mathrm{Unif}([p_e(H), (1+\beta)p_e(H)])$$

and assign the weight $1/\widetilde{p}_e$ to g_e. If e is not sampled, its weight is set to 0. We denote by $\tilde{H}$ the resulting randomly reweighted subgraph, and by H' the original (unperturbed) sparsifier that uses weights $1/p_e(H)$. Note that $\widetilde{p}_e$ is used solely as a weight parameter and is not required to lie in $[0, 1]$.

Preservation of approximation guarantee. By construction, whenever e is selected, we have

$$\frac{1}{(1+\beta)p_e(H)} \le \frac{1}{\widetilde{p}_e} \le \frac{1}{p_e(H)}.$$

Therefore, for every cut $A \subseteq U$, we have

$$\frac{1}{1+\beta}\,\mathrm{cut}_{H'}(A) \le \mathrm{cut}_{\tilde{H}}(A) \le \mathrm{cut}_{H'}(A).$$

Combining this inequality with the fact that H' is a $(1 \pm \xi)$-cut sparsifier of H, we obtain

$$\frac{1-\xi}{1+\beta}\,\mathrm{cut}_H(A) \le \mathrm{cut}_{\tilde{H}}(A) \le (1+\xi)\,\mathrm{cut}_H(A) \qquad \text{for all } A \subseteq U.$$

Hence $\tilde{H}$ is a $(1 \pm \varepsilon)$-cut sparsifier, and its exact anchored minimizer is a $(1+\varepsilon)$-approximate minimizer of the original objective by Remark K.2.

K.4 Bounding average sensitivity

We show that, given a submodular hypergraph $H = (U, E, \{g_e\}_{e\in E})$ with $|E| \ge 1$, the law of the output of the $(1 \pm \varepsilon)$-sparsification algorithm, equipped with the above weight perturbation, satisfies

$$\frac{1}{|E|}\sum_{e\in E} \mathrm{TV}(\mathcal{L}_\varepsilon(H), \mathcal{L}_\varepsilon(H-e)) = O\left(\frac{\varepsilon^{-3}(|U|^3 + |U|^2\log T)}{|E|}\right),$$

thereby proving Proposition 13.

Decomposition into independent coordinates. Let $Z = (Z_f)_{f\in E} \in \{0,1\}^E$ denote the random subset indicator vector, where $Z_f = 1$ iff $f \in E$ is sampled, and let $W = (W_f)_{f\in E} \in \mathbb{R}^E_{\ge 0}$ denote the resulting random weight vector (with $W_f = 0$ when $Z_f = 0$). Note that specifying these random variables determines the output sparsifier. Fix any $e \in E$, and let P and Q be the joint distributions of (Z, W) produced from H and $H - e$, respectively. We view both P and Q as probability measures on the common space $\{0,1\}^E \times \mathbb{R}^E_{\ge 0}$ by identifying the e-th coordinate of the sampling probability vector under $H - e$ with $p_e(H - e) = 0$. By the procedure of the above sparsification algorithm, for each $f \in E$, the random pair (Z_f, W_f) is generated independently of all the other coordinates, both under P and under Q. Therefore, P and Q admit the product decompositions

$$P = \bigotimes_{f\in E} P^{(f)} \qquad \text{and} \qquad Q = \bigotimes_{f\in E} Q^{(f)},$$

where $P^{(f)}$ and $Q^{(f)}$ are probability measures on $\{0,1\} \times \mathbb{R}_{\ge 0}$ corresponding to hyperedge f. We let $Q^{(e)} = \delta_{(0,0)}$, where $\delta_{(0,0)}$ is the Dirac measure at $(0, 0)$. By applying Lemma H.2, we have

$$\mathrm{TV}(P, Q) \le \sum_{f\in E} \mathrm{TV}\left(P^{(f)}, Q^{(f)}\right). \tag{11}$$

Below, we analyze the total variation distance for each coordinate f.

One-coordinate comparison. For $\beta > 0$ and $p \in [0,1]$, let $P_{p,\beta}$ be the law of (Z, W) with $Z \sim \mathrm{Bernoulli}(p)$ and $W = 0$ when $Z = 0$ holds; for $p > 0$, conditional on $Z = 1$, set $W = 1/\widetilde{p}$, where $\widetilde{p} \sim \mathrm{Unif}([p, (1+\beta)p])$. In particular, we have $P_{0,\beta} = \delta_{(0,0)}$. For $p, q \in (0,1]$ and $\beta, \gamma > 0$, Lemma H.3 and data processing under $x \mapsto 1/x$ give

$$\mathrm{TV}(P_{p,\beta}, P_{q,\gamma}) \le |p-q| + \min\{p,q\}\,\mathrm{TV}\big(\mathrm{Unif}([p,(1+\beta)p]), \mathrm{Unif}([q,(1+\gamma)q])\big). \tag{12}$$

For $p, q \in [0,1]$ with $p = 0$ or $q = 0$, one of the two laws is concentrated at $(0,0)$, and the other assigns probability $|p-q|$ to $Z = 1$. Therefore, we have $\mathrm{TV}(P_{p,\beta}, P_{q,\beta}) = |p-q|$. Here, we have $P^{(f)} = P_{p_f(H),\varepsilon/2}$ and $Q^{(f)} = P_{p_f(H-e),\varepsilon/2}$. For positive sampling probabilities, Lemma H.4 bounds the conditional total variation distance by $\frac{2+\varepsilon}{\varepsilon}|1 - p_f(H-e)/p_f(H)|$. Multiplying by $\min\{p_f(H), p_f(H-e)\} \le p_f(H)$ in (12), and including the zero-probability case, yields

$$\mathrm{TV}(P^{(f)}, Q^{(f)}) \le \left(1 + \frac{2+\varepsilon}{\varepsilon}\right)|p_f(H) - p_f(H-e)| \le \frac{4}{\varepsilon}|p_f(H) - p_f(H-e)|,$$

where the last inequality uses $\varepsilon \le 1$.

Bounding total variation over all hyperedges. Combining the above bound with (11) gives

$$\mathrm{TV}(P, Q) \le \frac{4}{\varepsilon}\sum_{f \in E}|p_f(H) - p_f(H-e)| = \frac{4}{\varepsilon}\left(\sum_{f \in E,\, f \ne e}|p_f(H) - p_f(H-e)| + p_e(H)\right).$$

By the definition of $p_f(H) = \min\{1,\ M s_f(H)\}$ and Lemmas K.3 and K.4, we have

$$\sum_{f \in E,\, f \ne e}|p_f(H) - p_f(H-e)| + p_e(H) \le M\left(\sum_{f \in E,\, f \ne e}|s_f(H) - s_f(H-e)| + s_e(H)\right) \le 2M s_e(H).$$

Therefore, taking the average over $e \in E$ and using (10) and $M = \Theta(\varepsilon^{-2}(|U| + \log T))$, we obtain

$$\begin{aligned}\frac{1}{|E|}\sum_{e \in E}\mathrm{TV}(\mathcal{L}_\varepsilon(H), \mathcal{L}_\varepsilon(H-e)) &\le \frac{8M}{\varepsilon|E|}\sum_{e \in E} s_e(H) \\ &\le \frac{8M}{\varepsilon|E|}(|U|^2 + 1) = O\left(\frac{\varepsilon^{-3}}{|E|}\left(|U|^3 + |U|^2 \log T\right)\right).\end{aligned}$$

This completes the proof of Proposition 13, and hence $\varphi(\varepsilon) = C_{\mathrm{sub}}\varepsilon^{-3}(n^3 + n^2 \log T)$.

K.5 Checking Assumption 9

We next verify that the above sparsifier construction, equipped with the weight perturbation, satisfies Assumption 9 with $C_{\mathrm{pc}} = O(1)$.

Proposition K.5. *Let $H = (U, E, \{g_e\}_{e \in E})$ be a submodular hypergraph, and let $\mathcal{L}_\varepsilon(H)$ be the law of the perturbed sparsifier above with parameter $\varepsilon \in (0,1]$. For $0 < \varepsilon \le \eta \le 1$, it holds that*

$$\mathrm{TV}(\mathcal{L}_\varepsilon(H), \mathcal{L}_\eta(H)) = O\left((|U|^3 + |U|^2 \log T)\varepsilon^{-3}\frac{\eta - \varepsilon}{\varepsilon}\right).$$

Consequently, for the anchored lifted instance in Appendix K.1 built from the original n-element ground set, the sparsifier-based offline algorithm family satisfies Assumption 9 with $C_{\mathrm{pc}} = O(1)$.

Proof. For each parameter $\alpha \in (0,1]$, represent the output of the sparsifier construction by the vector $((Z_f^\alpha, W_f^\alpha))_{f \in E}$, where Z_f^α indicates whether f is sampled and $W_f^\alpha = 0$ when $Z_f^\alpha = 0$. Conditional on $Z_f^\alpha = 1$ and $p_f^\alpha(H) > 0$, the algorithm draws $\widetilde{p}_f^\alpha$ uniformly from $[p_f^\alpha(H), (1 + \alpha/2)p_f^\alpha(H)]$ and sets $W_f^\alpha = 1/\widetilde{p}_f^\alpha$, where

$$p_f^\alpha(H) = \min\{1, M(\alpha)s_f(H)\}, \qquad M(\alpha) = c_{\mathrm{sp}}\alpha^{-2}(|U| + \log(2T)),$$

for a fixed absolute constant $c_{\mathrm{sp}} > 0$. Since $x \mapsto \min\{1, x\}$ is 1-Lipschitz, $\sum_f s_f(H) \le |U|^2 + 1$, and $|M(\varepsilon) - M(\eta)| = O(M(\varepsilon)(\eta - \varepsilon)/\varepsilon)$ for $0 < \varepsilon \le \eta \le 1$, we have

$$\sum_{f \in E} |p_f^\varepsilon(H) - p_f^\eta(H)| = O\left((|U|^3 + |U|^2 \log T)\varepsilon^{-2} \frac{\eta - \varepsilon}{\varepsilon}\right). \tag{13}$$

Also, it holds that

$$\sum_{f \in E} p_f^\varepsilon(H) \le M(\varepsilon) \sum_{f \in E} s_f(H) = O\big((|U|^3 + |U|^2 \log T)\varepsilon^{-2}\big). \tag{14}$$

For each $f \in E$ and $\alpha \in (0, 1]$, write $P_f^\alpha := P_{p_f^\alpha(H), \alpha/2}$. If $p_f^\varepsilon(H) > 0$ holds, then $s_f(H) > 0$ and $p_f^\eta(H) > 0$ also hold. In this case, Lemma I.5 gives

$$\begin{aligned}
&\mathrm{TV}\Big(\mathrm{Unif}([p_f^\varepsilon(H), (1 + \varepsilon/2)p_f^\varepsilon(H)]), \mathrm{Unif}([p_f^\eta(H), (1 + \eta/2)p_f^\eta(H)])\Big) \\
&= O\left(\frac{1}{\varepsilon}\left|1 - \frac{p_f^\eta(H)}{p_f^\varepsilon(H)}\right| + \frac{\eta - \varepsilon}{\varepsilon}\right).
\end{aligned}$$

Substituting this bound into (12) and using $\min\{p_f^\varepsilon(H), p_f^\eta(H)\} \le p_f^\varepsilon(H)$ yield

$$\mathrm{TV}\Big(P_f^\varepsilon, P_f^\eta\Big) = O\left(\frac{1}{\varepsilon}|p_f^\varepsilon(H) - p_f^\eta(H)| + p_f^\varepsilon(H)\frac{\eta - \varepsilon}{\varepsilon}\right).$$

If $p_f^\varepsilon(H) = 0$, then we have $s_f(H) = 0$ and therefore $p_f^\eta(H) = 0$, so the same bound is trivial. Since the coordinates are sampled independently under each parameter, Lemma H.2 gives

$$\begin{aligned}
\mathrm{TV}(\mathcal{L}_\varepsilon(H), \mathcal{L}_\eta(H)) &\le \sum_{f \in E} \mathrm{TV}\Big(P_f^\varepsilon, P_f^\eta\Big) \\
&= O\left(\frac{1}{\varepsilon}\sum_{f \in E}|p_f^\varepsilon(H) - p_f^\eta(H)| + \frac{\eta - \varepsilon}{\varepsilon}\sum_{f \in E} p_f^\varepsilon(H)\right) \\
&= O\left((|U|^3 + |U|^2 \log T)\varepsilon^{-3}\frac{\eta - \varepsilon}{\varepsilon}\right),
\end{aligned}$$

where the last step uses (13) and (14), together with $\varepsilon \le 1$.

To establish the condition in Assumption 9, let X be a multiset of losses and let H_X^+ be its anchored lifted hypergraph. We have

$$\mathrm{OPT}(X) = \min_{S \subseteq V} \mathrm{cut}_{H_X^+}(S \cup \{r\}) = \min_{S \subseteq V} \sum_{\ell \in X} \ell(S),$$

and set $\varepsilon_X = \varepsilon_{\min}(\mathrm{OPT}(X) + \lambda)$. For each $x \in X$, let e_x be the corresponding lifted hyperedge and write $Y = X \setminus x$, so that $H_Y^+ = H_X^+ - e_x$ and $\varepsilon_Y = \varepsilon_{\min}(\mathrm{OPT}(Y) + \lambda)$. Since the final decision is obtained from the sparsifier by deterministic minimization, the data processing inequality gives

$$\mathrm{TV}(\mathcal{A}_{\varepsilon_X}(Y), \mathcal{A}_{\varepsilon_Y}(Y)) \le \mathrm{TV}\big(\mathcal{L}_{\varepsilon_X}(H_Y^+), \mathcal{L}_{\varepsilon_Y}(H_Y^+)\big).$$

Since all losses take values in $[0, 1]$, we have

$$0 \le \mathrm{OPT}(X) - \mathrm{OPT}(Y) \le 1.$$

Indeed, monotonicity follows from non-negativity, and the upper bound follows by evaluating the added loss at an optimizer for H_Y^+. Applying the fixed-hypergraph bound above to H_Y^+ with $|U| = |V^+| = n+1$, $\varepsilon = \varepsilon_X$, and $\eta = \varepsilon_Y$, and using Lemma J.1 with $\varphi(\varepsilon) = C_{\mathrm{sub}}\varepsilon^{-3}(n^3 + n^2 \log T)$, yield

$$\begin{aligned}
\mathrm{TV}(\mathcal{A}_{\varepsilon_X}(Y), \mathcal{A}_{\varepsilon_Y}(Y)) &= O\left((n^3 + n^2 \log T)\varepsilon_X^{-3}\frac{\varepsilon_Y - \varepsilon_X}{\varepsilon_X}\right) \\
&= O(\varepsilon_X(\mathrm{OPT}(X) - \mathrm{OPT}(Y))).
\end{aligned}$$

Averaging this inequality over $x \in X$ gives the condition in Assumption 9 for the original submodular function minimization instance, with $C_{\mathrm{pc}} = O(1)$. □

L Offline algorithm for ℓ_1 regression

We present the details of the offline algorithm for ℓ_1 regression used in Section 6. We use the ℓ_1 regression algorithm of Parulekar et al. [46] based on Lewis weight sampling. We analyze the offline algorithm on a dataset $X = \{(a_i^{\mathrm{feat}}, b_i)\}_{i=1}^{n_0} \subset \mathbb{R}^{d_{\mathrm{feat}}} \times \mathbb{R}$ of n_0 labeled feature–response pairs, where a_i^{feat} is the feature vector and b_i is the response for each $i \in [n_0]$. Assume the feature dimension satisfies $d_{\mathrm{feat}} \geq 1$, set $d := d_{\mathrm{feat}} + 1 \leq 2d_{\mathrm{feat}}$ for the augmented sampling matrix, defined below, and put $\delta_T := (2T)^{-1}$. Let the feature matrix $A^{\mathrm{feat}} \in \mathbb{R}^{n_0 \times d_{\mathrm{feat}}}$ have rows $(a_i^{\mathrm{feat}})^\top$, and let $b := (b_1, \ldots, b_{n_0})^\top$. Define the augmented sampling matrix

$$A := [A^{\mathrm{feat}}, -b] \in \mathbb{R}^{n_0 \times d}.$$

We write $a_i^\top = ((a_i^{\mathrm{feat}})^\top, -b_i)$ for the i-th row of this augmented matrix, so that $|\langle a_i^{\mathrm{feat}}, \theta\rangle - b_i| = |\langle a_i, (\theta, 1)\rangle|$. Rows with $a_i = 0$, equivalently $(a_i^{\mathrm{feat}}, b_i) = (0, 0)$, have zero loss for all θ and may be discarded; rows with $a_i^{\mathrm{feat}} = 0$ and $b_i \neq 0$ remain non-zero augmented rows. After discarding zero augmented rows, relabel the remaining rows and write $n \leq n_0$ for their number. Deleting any discarded row leaves the output law unchanged, and hence the discarded rows contribute zero to the sum of deletion total-variation distances. Consequently, dividing the sum bound proved below by the original size n_0 gives the desired average-sensitivity bound in Proposition L.7. If no row remains, the routine returns the fixed minimizer of the zero objective. Throughout the rest of this appendix, A denotes this augmented matrix, after zero augmented rows have been discarded when necessary, or a row-deleted version of it, and d denotes its column dimension. Since $d = d_{\mathrm{feat}} + 1 \leq 2d_{\mathrm{feat}}$, the bounds stated below imply the rates in Section 6, after replacing d_{feat} with d used there for the feature dimension and adjusting absolute constants.

Remark L.1. Dong and Yoshida [19] also studied online regression, but their argument does not directly apply to the batch-to-online framework. Specifically, their offline approximation guarantee is given in terms of the (unsquared) ℓ_2 error based on an ℓ_2-norm subspace embedding, but this error measure is not additive over rounds and therefore cannot be plugged into the batch-to-online framework. If one instead considers the squared ℓ_2 error, the induced loss becomes exp-concave, and the standard online Newton method yields a stronger guarantee.

L.1 ℓ_1 Lewis weights and sampling algorithm

We define ℓ_1 Lewis weights and introduce useful properties. In accordance with the zero-row discard and relabeling convention above, for a matrix $A \in \mathbb{R}^{n \times d}$ with non-zero rows, the ℓ_1 Lewis weights $w(A) = (w_1(A), \ldots, w_n(A))$ are defined as the unique positive solution to the system of equations

$$(w_i(A))^2 = a_i^\top \left(A^\top W(A)^{-1} A\right)^\dagger a_i, \qquad i \in [n],$$

where we define $W(A) = \mathrm{diag}(w_1(A), \ldots, w_n(A))$ and $M^\dagger$ denotes the Moore–Penrose pseudoinverse of a matrix M. For a matrix with no rows, let $w(A)$ be the empty vector. For symmetric matrices $M, N \in \mathbb{R}^{d \times d}$, write $M \succeq N$, equivalently $N \preceq M$, when $v^\top (M - N) v \geq 0$ for every $v \in \mathbb{R}^d$. Below are basic properties of ℓ_1 Lewis weights (cf. [17]). We include a short proof for completeness.

Lemma L.2 (Basic properties of ℓ_1 Lewis weights)**.** *Let $A \in \mathbb{R}^{n \times d}$ be a matrix with non-zero rows. The ℓ_1 Lewis weights $w(A)$ satisfy the following properties:*

1. *Weights are positive: $w_i(A) > 0$ for all $i \in [n]$.*
2. *Bound on the sum of weights: $\displaystyle\sum_{i=1}^{n} w_i(A) = \mathrm{rank}(A) =: r \leq d$.*
3. *Monotonicity: if A' is obtained from A by appending finitely many non-zero rows, then for every original row $j \in [n]$, we have*
$$w_j(A') \leq w_j(A).$$

Proof. If $n = 0$, the first and third claims are vacuous, and the second holds because the empty sum and $\mathrm{rank}(A)$ both equal zero. We may therefore assume $n \geq 1$. Let $w = w(A)$ and $W = W(A)$ for

brevity. We use the fixed-point characterization of ℓ_p Lewis weights: for $p > 0$ and full-column-rank A, these are the positive weights $u \in \mathbb{R}^n_{>0}$ satisfying $u_i^{2/p} = a_i^\top (A^\top \operatorname{diag}(u)^{1-2/p} A)^{-1} a_i$ for every $i \in [n]$ [17, Definition 2.2]. For $p = 1 < 4$, Cohen and Peng [17, Corollary 3.4] gives existence and uniqueness on full-column-rank matrices. For arbitrary-rank A, let V have orthonormal columns spanning $\operatorname{range}(A^\top)$ and put $\widehat{A} \coloneqq AV$. Then $\widehat{A}$ has full column rank and non-zero rows. Since $A = \widehat{A} V^\top$ and $V^\top V = I_r$, for every positive diagonal matrix $D \in \mathbb{R}^{n \times n}$ and every $i \in [n]$, we have

$$a_i^\top (A^\top D^{-1} A)^\dagger a_i = a_i^\top V (\widehat{A}^\top D^{-1} \widehat{A})^{-1} V^\top a_i = \widehat{a}_i^\top (\widehat{A}^\top D^{-1} \widehat{A})^{-1} \widehat{a}_i,$$

where $\widehat{a}_i \coloneqq V^\top a_i$. Thus, A and $\widehat{A}$ have the same positive solutions to the Lewis-weight equations, proving the first claim.

For the second claim, let $B \coloneqq W^{-1/2} A$ and $M \coloneqq A^\top W^{-1} A = B^\top B$. Let $b_i^\top$ denote the i-th row of B, i.e., $b_i = w_i^{-1/2} a_i$. Then, from $w_i^2 = a_i^\top M^\dagger a_i$, we have

$$\tau_i(B) \coloneqq b_i^\top (B^\top B)^\dagger b_i = w_i^{-1} a_i^\top M^\dagger a_i = w_i^{-1} w_i^2 = w_i.$$

That is, w_i coincides with the statistical leverage score of the reweighted matrix B. It holds that

$$\sum_{i=1}^n w_i = \sum_{i=1}^n \tau_i(B) = \operatorname{tr}\big(B (B^\top B)^\dagger B^\top\big) = \operatorname{tr}\big((B^\top B)(B^\top B)^\dagger\big) = \operatorname{rank}(B).$$

Finally, since $W^{-1/2}$ is an invertible diagonal matrix, $\operatorname{rank}(B) = \operatorname{rank}(A) = r$ holds, proving the claim.

For the third claim, first suppose that one non-zero row is added, and write w and w' for the weights before and after the addition. Let

$$\rho \coloneqq \max_{j \in [n]} w'_j / w_j, \qquad G \coloneqq A^\top W^{-1} A, \qquad \text{and} \qquad H \coloneqq (A')^\top \operatorname{diag}(w')^{-1} A'.$$

Since $w'_j \le \rho w_j$ for every original row, we have

$$H \succeq \sum_{j=1}^n \frac{a_j a_j^\top}{w'_j} \succeq \rho^{-1} \sum_{j=1}^n \frac{a_j a_j^\top}{w_j} = \rho^{-1} G. \tag{15}$$

For each $j \in [n]$, we have $a_j \in \operatorname{range}(G)$. Applying the range-restricted pseudoinverse inequality to (15) and using the fixed-point equations gives

$$(w'_j)^2 = a_j^\top H^\dagger a_j \le \rho\, a_j^\top G^\dagger a_j = \rho w_j^2 \qquad (j \in [n]).$$

At a row attaining the maximum ratio, this gives $\rho^2 w_j^2 \le \rho w_j^2$, and hence $\rho \le 1$. This is the maximum-ratio argument of Cohen and Peng [16, Appendix A, proof of Lemma 5.5], with the inverse comparison replaced by its range-restricted Moore–Penrose form. Repeating the one-row conclusion for each added row completes the claim. □

We first analyze sampling with exact Lewis weights. Their deletion monotonicity controls the change in the sampling distribution, while randomized rounding and weight perturbation control the effect of changing the target accuracy. Deterministic polynomial-time approximations to these weights retain the regression approximation and both stability guarantees; Appendix L.5 gives the computation and its analysis. For a retained matrix A with at least one row, define

$$p_i(A) \coloneqq \frac{w_i(A)}{\sum_{j=1}^n w_j(A)}, \qquad i \in [n].$$

For an input with no non-zero augmented row, set $p(A) = 0$ on its original row index set.

Let $S = ((u_j, v_j, \omega_j))_{j=1}^m$ be a finite ordered weighted sample, where $m \ge 0$ is an integer. For each $j \in [m]$, the entry consists of a feature vector $u_j \in \mathbb{R}^{d_{\text{feat}}}$, a response $v_j \in \mathbb{R}$, and a weight $\omega_j \ge 0$. Let $\theta^\star(S)$ denote the output of the exact optimization routine assumed in Section 6, with its fixed tie-breaking rule; we retain the sample order because the tie-breaking rule may depend on it. The output satisfies

$$\theta^\star(S) \in \arg\min_{\theta \in \Theta} \sum_{j=1}^m \omega_j |\langle u_j, \theta \rangle - v_j|.$$

For the empty sample, the routine returns a fixed point $\theta^\star(\emptyset) \in \Theta$. For each fixed sample length, we assume that the optimizer $S \mapsto \theta^\star(S)$ is Borel-measurable, in accordance with the measurability requirement in Section 2.

Offline ℓ_1 regression procedure with exact Lewis weights. Given A, the feasible set Θ, a target accuracy $\varepsilon \in (0, 1]$, and a mean sample size $\mu \geq 1$, discard the zero augmented rows. If none remain, return $\theta^\star(\emptyset)$. Otherwise, perform the following steps.

1. Form the sampling distribution $p(A)$ from the exact Lewis weights.
2. Draw $N_\varepsilon \sim R_\mu$ and, conditional on N_ε, draw $I_1, \ldots, I_{N_\varepsilon}$ independently from $p(A)$.
3. For each sampled occurrence $h \in [N_\varepsilon]$, independently draw
$$\widetilde{p}_h \sim \mathrm{Unif}([p_{I_h}(A), (1 + \varepsilon/2)p_{I_h}(A)]).$$
4. Form the ordered weighted sample
$$\widehat{S} \coloneqq \left(\left(a_{I_h}^{\mathrm{feat}}, b_{I_h}, \frac{1}{N_\varepsilon \widetilde{p}_h}\right)\right)_{h=1}^{N_\varepsilon}$$
and return $\theta^\star(\widehat{S})$. The weighted objective for $\widehat{S}$ is
$$\widehat{L}(\theta) = \sum_{h=1}^{N_\varepsilon} \frac{|\langle a_{I_h}, (\theta, 1)\rangle|}{N_\varepsilon \widetilde{p}_h}.$$

Let $\mathcal{L}_\varepsilon^{\mathrm{ex}}(A)$ denote the output law of the procedure using exact Lewis weights, with the mean μ_ε specified below.

Approximation guarantee. We analyze how sampling changes the ℓ_1 loss. To cover both exact and approximately computed Lewis weights, we allow any positive probability vector q satisfying $q_i \geq \mathrm{e}^{-2} p_i(A)$ for every $i \in [n]$. The factor e^{-2} accounts for approximation errors in both the individual weights and their normalizing sum. This condition holds for $p(A)$ and for the normalized approximate weights we compute; see Appendix L.5 for the computation and the proof of this bound. Fix a universal coefficient $c_{\ell_1} \geq 1$, whose value is specified in the proof of Lemma L.4, take $\mu = \mu_\varepsilon$, and set
$$\mu_\varepsilon = c_{\ell_1} d \varepsilon^{-2} \log\left(\mathrm{e} + \frac{2dT}{\varepsilon}\right) + 2.$$

Put $N_\varepsilon^- \coloneqq \lfloor \mu_\varepsilon \rfloor$ and $N_\varepsilon^+ \coloneqq \lceil \mu_\varepsilon \rceil$. For a positive probability vector $q \in \mathbb{R}^n$ and $N \in \{N_\varepsilon^-, N_\varepsilon^+\}$, draw $I_1, \ldots, I_N$ independently from q and let $\mathcal{E}_A(\varepsilon, N; q)$ be the event that
$$(1 - \varepsilon/12)\|Az\|_1 \leq \sum_{k=1}^{N} \frac{|\langle a_{I_k}, z\rangle|}{N q_{I_k}} \leq (1 + \varepsilon/12)\|Az\|_1$$
holds simultaneously for every $z \in \mathbb{R}^d$.

We first consider matrices of full column rank and fix the number of sampled rows. We give sufficient conditions for the approximation above to hold, allowing each row to be sampled with probability at least a fraction $\kappa \in (0, 1]$ of its normalized Lewis weight.

Lemma L.3. *There are universal constants $C_{\mathrm{emb}}, C_0 \geq 1$ with the following property. Let $B \in \mathbb{R}^{m \times r}$ have full column rank and no zero rows, write $b_i^\top$ for its rows and $w_i(B)$ for its exact ℓ_1 Lewis weights, and let $q \in \mathbb{R}^m_{>0}$ satisfy $\sum_i q_i = 1$. Suppose that $N \in \mathbb{Z}_{>0}$, $0 < \kappa, \xi \leq 1$, $0 < \delta \leq 1/2$, and*
$$q_i \geq \frac{\kappa w_i(B)}{r} \quad (i \in [m]), \qquad \frac{\kappa N}{r} \geq C_{\mathrm{emb}} \xi^{-2} \log\left(\frac{2\mathrm{e} C_0 (N + r^2)}{\delta}\right). \tag{16}$$
If $I_1, \ldots, I_N$ are independent samples from q, then, with probability at least $1 - \delta$, we have
$$(1 - \xi)\|Bz\|_1 \leq \sum_{k=1}^{N} \frac{|\langle b_{I_k}, z\rangle|}{N q_{I_k}} \leq (1 + \xi)\|Bz\|_1 \qquad \textit{for every } z \in \mathbb{R}^r.$$

Proof. Put $G \coloneqq \kappa N/r$ and $\gamma_i \coloneqq Nq_i$. The $p = 1$ case of Cohen and Peng [16, Lemma B.1] gives universal constants $C_{\mathrm{row}}, C_{\mathrm{obj}}, C_{\mathrm{wt}} \geq 1$ and an auxiliary matrix B_{aux} having at most $C_{\mathrm{row}} r^2$ rows for which, for every $z \in \mathbb{R}^r$, we have

$$B^\top W(B)^{-1} B \preceq B_{\mathrm{aux}}^\top W(B_{\mathrm{aux}})^{-1} B_{\mathrm{aux}}, \qquad \|B_{\mathrm{aux}} z\|_1 \leq C_{\mathrm{obj}} \|Bz\|_1,$$

and every row of B_{aux} has Lewis weight at most C_{wt}/r. To obtain rows whose Lewis weights are at most $1/G$, let $L_{\mathrm{split}} \coloneqq \lceil C_{\mathrm{wt}} G/r \rceil$ and replace each row $a^\top$ of B_{aux}, with Lewis weight w, by L_{split} copies of $a^\top/L_{\mathrm{split}}$. Assigning weight w/L_{split} to each copy preserves the matrix $B_{\mathrm{aux}}^\top W(B_{\mathrm{aux}})^{-1} B_{\mathrm{aux}}$, since we have

$$L_{\mathrm{split}} \frac{(a/L_{\mathrm{split}})(a/L_{\mathrm{split}})^\top}{w/L_{\mathrm{split}}} = \frac{aa^\top}{w}.$$

The fixed-point equations and uniqueness therefore show that each copy has Lewis weight w/L_{split}. The bound $w/L_{\mathrm{split}} \leq C_{\mathrm{wt}}/(rL_{\mathrm{split}}) \leq 1/G$ holds. Replacing the rows in this way also preserves $\|B_{\mathrm{aux}} z\|_1$ for every $z \in \mathbb{R}^r$, so both preceding inequalities remain valid after the replacements.

Write $I \coloneqq (I_1, \ldots, I_N)$ for the sampled labels. Fix their values and append the N rows $b_{I_k}^\top/\gamma_{I_k}$, $k \in [N]$, to the matrix formed by these copies. Call the result B_I and write $H_I \coloneqq B_I^\top W(B_I)^{-1} B_I$. By Lemma L.2 (3), appending rows cannot increase the weights of these copies. The row replacements preserve $B_{\mathrm{aux}}^\top W(B_{\mathrm{aux}})^{-1} B_{\mathrm{aux}}$, so we have

$$H_I \succeq B_{\mathrm{aux}}^\top W(B_{\mathrm{aux}})^{-1} B_{\mathrm{aux}} \succeq B^\top W(B)^{-1} B \succ 0. \tag{17}$$

For each sampled occurrence $k \in [N]$, let $w_{k,I}$ denote its Lewis weight in B_I. Taking inverses reverses the matrix order in (17), so the fixed-point equations give

$$w_{k,I}^2 = \frac{b_{I_k}^\top H_I^{-1} b_{I_k}}{\gamma_{I_k}^2} \leq \frac{b_{I_k}^\top (B^\top W(B)^{-1} B)^{-1} b_{I_k}}{\gamma_{I_k}^2} = \frac{w_{I_k}(B)^2}{\gamma_{I_k}^2}.$$

Hence $w_{k,I} \leq w_{I_k}(B)/\gamma_{I_k} \leq 1/G$ by (16). The copies of the rows of B_{aux} also have Lewis weights at most $1/G$ in B_I by Lemma L.2 (3). The number of rows in B_I, denoted by m_I, does not depend on the sampled labels and satisfies

$$m_I \leq N + C_{\mathrm{row}} r^2 L_{\mathrm{split}} \leq C_0 (N + r^2), \tag{18}$$

where we fix $C_0 \geq 1 + C_{\mathrm{row}} + C_{\mathrm{row}} C_{\mathrm{wt}}$ and use $L_{\mathrm{split}} \leq C_{\mathrm{wt}} G/r + 1$, $rG = \kappa N$, and $\kappa \leq 1$.

We next use independent random signs to control the error caused by sampling. Let $s \coloneqq \lceil \log(2m_I/\delta) \rceil$ and $\xi_0 \coloneqq \xi/(4(1 + C_{\mathrm{obj}}))$. Write $(b_j^I)^\top$ for row $j \in [m_I]$ of B_I, and draw $\sigma_1, \ldots, \sigma_{m_I}$ independently and uniformly from $\{-1, 1\}$. Define

$$\mathcal{R}_I(\sigma) \coloneqq \sup_{\|B_I z\|_1 = 1} \left| \sum_{j=1}^{m_I} \sigma_j \left| \langle b_j^I, z \rangle \right| \right|.$$

By Cohen and Peng [17, Lemma 8.4], there is a universal constant $C_{\mathrm{mom}} \geq 1$ for which the s-th moment satisfies

$$\mathbb{E}_\sigma[\mathcal{R}_I(\sigma)^s] \leq 2m_I \left(\frac{C_{\mathrm{mom}} s}{G} \right)^{s/2} \leq \xi_0^s \delta. \tag{19}$$

Indeed, the definition of s gives $2m_I \leq \delta \mathrm{e}^s$, while (18) gives

$$s \leq \log \left(\frac{2\mathrm{e} C_0 (N + r^2)}{\delta} \right).$$

Thus the second inequality in (19) follows from (16) once $C_{\mathrm{emb}} \geq 16\mathrm{e}^2 C_{\mathrm{mom}} (1 + C_{\mathrm{obj}})^2$.

We now use (19) to bound the error caused by sampling. Define

$$\Delta(I) \coloneqq \sup_{\|Bz\|_1 = 1} \left| \sum_{k=1}^{N} \frac{|\langle b_{I_k}, z \rangle|}{\gamma_{I_k}} - 1 \right|.$$

We apply the symmetrization argument in Cohen and Peng [16, Appendix B, proof of Lemma 7.4] to the N independent draws and the matrix obtained from B_{aux} by the row replacements. For every $z \in \mathbb{R}^r$, we have $\|B_I z\|_1 \le (1 + C_{\mathrm{obj}} + \Delta(I))\|Bz\|_1$. Combining this inequality with (19) in the symmetrization argument gives

$$\mathbb{E}_I[\Delta(I)^s] \le 2^{2s-1}\xi_0^s\delta((1 + C_{\mathrm{obj}})^s + \mathbb{E}_I[\Delta(I)^s]).$$

The coefficient of the final expectation is $(4\xi_0)^s\delta/2 \le 1/2$. Moving that term to the left gives $\mathbb{E}_I[\Delta(I)^s] \le \xi^s\delta$, and Markov's inequality gives $\Pr[\Delta(I) > \xi] \le \delta$. □

We next express the rows of A in an orthonormal basis of its row space and apply Lemma L.3. We verify its requirement on the number of sampled rows separately for N_ε^- and N_ε^+.

Lemma L.4. *For the universal coefficient c_{ℓ_1} specified in (22), every retained augmented matrix A obtained from a labeled prefix of size $1 \le t \le T$ and having at least one row satisfies the following: for every positive probability vector $q \in \mathbb{R}^n$ with $q_i \ge \mathrm{e}^{-2}w_i(A)/\operatorname{rank}(A)$ for all $i \in [n]$, we have*

$$\Pr[\mathcal{E}_A(\varepsilon, N; q)] \ge 1 - \delta_T \qquad \text{for each } N \in \{N_\varepsilon^-, N_\varepsilon^+\} \tag{20}$$

for every $\varepsilon \in (0, 1]$.

Proof. Fix constants C_{emb}, C_0 for which Lemma L.3 holds. Let $r := \operatorname{rank}(A)$. Since A has a non-zero row, we have $r \ge 1$. Let V have orthonormal columns spanning $\operatorname{range}(A^\top)$, and put $\widehat{A} := AV$, as in the proof of Lemma L.2. Then $A = \widehat{A}V^\top$ holds, and $z \mapsto V^\top z$ maps $\mathbb{R}^d$ onto $\mathbb{R}^r$. Thus replacing A by $\widehat{A}$ preserves the exact Lewis weights and the event $\mathcal{E}_A(\varepsilon, N; q)$. The assumed lower bound on q_i gives the first condition of (16) with $\kappa = \mathrm{e}^{-2}$.

Set $\xi := \varepsilon/12$, $\delta := \delta_T$, and

$$\Lambda_\varepsilon := \log\left(\mathrm{e} + \frac{2dT}{\varepsilon}\right).$$

The definitions of μ_ε, N_ε^-, and N_ε^+ give

$$c_{\ell_1} d\varepsilon^{-2}\Lambda_\varepsilon + 1 \le N_\varepsilon^- \le N_\varepsilon^+ \le c_{\ell_1} d\varepsilon^{-2}\Lambda_\varepsilon + 3. \tag{21}$$

For each $N \in \{N_\varepsilon^-, N_\varepsilon^+\}$, the upper bound in (21), $r \le d$, and $d, T, \varepsilon^{-1}, \Lambda_\varepsilon \ge 1$ imply

$$N + r^2 \le (c_{\ell_1} + 4)d^2\varepsilon^{-2}\Lambda_\varepsilon.$$

Since $d^2T\varepsilon^{-2} \le (dT/\varepsilon)^2$, $\log(dT/\varepsilon) \le \Lambda_\varepsilon$, and $\log\Lambda_\varepsilon \le \Lambda_\varepsilon$, we obtain

$$\log\left(\frac{2\mathrm{e}C_0(N + r^2)}{\delta_T}\right) \le (5 + \log(4\mathrm{e}C_0(c_{\ell_1} + 4)))\Lambda_\varepsilon.$$

We choose c_{ℓ_1} explicitly as follows. Let

$$A_0 := 144C_{\mathrm{emb}}\mathrm{e}^2, \quad B_0 := 5 + \log(4\mathrm{e}C_0), \quad H_0 := A_0 + B_0 + 1, \quad c_{\ell_1} := 16H_0^2. \tag{22}$$

The inequalities $\log(c_{\ell_1} + 4) \le 2\sqrt{c_{\ell_1} + 4} \le 10H_0$, $A_0 \le H_0$, and $B_0 \le H_0$ give

$$144C_{\mathrm{emb}}(5 + \log(4\mathrm{e}C_0(c_{\ell_1} + 4))) \le \mathrm{e}^{-2}c_{\ell_1}.$$

For either value of N, (21) and $r \le d$ therefore give

$$\frac{\mathrm{e}^{-2}N}{r} \ge \mathrm{e}^{-2}c_{\ell_1}\varepsilon^{-2}\Lambda_\varepsilon \ge C_{\mathrm{emb}}(\varepsilon/12)^{-2}\log\left(\frac{2\mathrm{e}C_0(N + r^2)}{\delta_T}\right).$$

Thus N_ε^- and N_ε^+ are positive and each satisfies (16). Applying Lemma L.3 separately with $N = N_\varepsilon^-$ and $N = N_\varepsilon^+$ proves (20). □

For the procedure using exact Lewis weights, take $q = p(A)$. Since $p_i(A) = w_i(A)/\operatorname{rank}(A)$ for every $i \in [n]$, the lower bound required by Lemma L.4 holds. By Lemma I.1, the randomly rounded count belongs to $\{N_\varepsilon^-, N_\varepsilon^+\}$ almost surely. Conditioning on its value and applying (20) gives failure probability at most δ_T. On the complementary event, Remark K.2 with $\xi = \varepsilon/12$ and $\beta = \varepsilon/2$ gives a $(1 + \varepsilon)$-approximate output. A size-n input has total loss at most n, and thus the failure event contributes at most $n\delta_T = n/(2T)$ to its expected loss.

L.2 Change of probability vectors under row deletion

To analyze the average sensitivity of the above procedure, we examine how the probability vector based on the Lewis weights changes when a row is deleted.

For each $i \in [n]$, let $A^{(-i)} \in \mathbb{R}^{(n-1)\times d}$ denote the matrix obtained from A by deleting the i-th row $a_i^\top$, and let $w(A^{(-i)})$ be the Lewis weight vector of $A^{(-i)}$. Here $w_j(A^{(-i)})$ denotes the weight of the surviving row $a_j^\top$, retaining its original label $j \in [n] \setminus \{i\}$. We define $w^{(-i)} \in \mathbb{R}^n$ by inserting a zero in the missing coordinate as follows:

$$w^{(-i)} \coloneqq (w_1^{(-i)}, \dots, w_n^{(-i)}) \quad \text{with} \quad w_i^{(-i)} \coloneqq 0 \quad \text{and} \quad w_j^{(-i)} \coloneqq w_j(A^{(-i)}) \quad \text{for } j \neq i.$$

Note that A can be obtained from $A^{(-i)}$ by adding the row $a_i^\top$. Thus, Lemma L.2 (3) implies

$$w_j(A) \leq w_j(A^{(-i)}) = w_j^{(-i)} \quad \text{for } j \neq i.$$

From Lemma L.2 (2), we also have $\sum_{j=1}^n w_j(A) = r$. Recall that the sampling distribution is

$$p(A) = (p_1(A), \dots, p_n(A)) \quad \text{where} \quad p_i(A) = \frac{w_i(A)}{r}.$$

Similarly, for $A^{(-i)}$, define $p(A^{(-i)})$ as the normalized Lewis weights of $A^{(-i)}$, extended by a zero in the i-th coordinate, i.e.,

$$p_j(A^{(-i)}) = \frac{w_j^{(-i)}}{r^{(-i)}} \quad \text{for } j \in [n],$$

where $r^{(-i)} = \sum_{j=1}^n w_j^{(-i)} = \operatorname{rank}(A^{(-i)}) \leq r$. If $r^{(-i)} = 0$, define $p(A^{(-i)})$ as the zero vector on the extended coordinate space. By using these, we obtain the following lemma on the average change of the exact probability vectors.

Lemma L.5. *For $n \geq 1$, we have*

$$\frac{1}{n}\sum_{i=1}^n \left\| p(A) - p(A^{(-i)}) \right\|_1 \leq \frac{2}{n}.$$

Proof. Write $w \coloneqq w(A)$. Fix $i \in [n]$. If $r^{(-i)} = 0$, then $p(A^{(-i)})$ is the zero vector. Moreover, $w_j(A) \leq w_j(A^{(-i)}) = 0$ holds for every $j \neq i$, so $r = \sum_j w_j(A) = w_i(A)$ holds and hence

$$\left\| p(A) - p(A^{(-i)}) \right\|_1 = 1 \leq \frac{2w_i}{r}.$$

Below, we assume $r^{(-i)} > 0$. Then, we have

$$\begin{aligned}
\left\| p(A) - p(A^{(-i)}) \right\|_1 &= \sum_{j=1}^n \left| \frac{w_j}{r} - \frac{w_j^{(-i)}}{r^{(-i)}} \right| \\
&= \frac{w_i}{r} + \sum_{j\neq i} \left(\frac{w_j^{(-i)}}{r^{(-i)}} - \frac{w_j}{r} \right) \\
&= \frac{w_i}{r} + \frac{\sum_{j\neq i} w_j^{(-i)}}{r^{(-i)}} - \frac{\sum_{j=1}^n w_j - w_i}{r} \\
&= \frac{w_i}{r} + \frac{r^{(-i)}}{r^{(-i)}} - \frac{r - w_i}{r} \\
&= \frac{2w_i}{r}.
\end{aligned}$$

Averaging over $i = 1, \dots, n$ and using $\sum_{i=1}^n w_i = r$ yield the claim. □

L.3 Bounding average sensitivity

We bound the average sensitivity of the procedure with exact Lewis weights. For the comparison below, we will express the outputs before and after row deletion as the same measurable function of the sampling transcripts. This allows us to bound the total variation distance between the output laws by that between the transcript laws. The following construction gives this representation.

Common measurable output map. Fix an integer $N \geq 1$. For a matrix A with rows indexed by $[n]$ and a sampling transcript $\tau = ((i_k, \widetilde{q}_k))_{k=1}^N \in [n]^N \times \mathbb{R}^N$, define, for each $k \in [N]$,

$$\omega_k(\tau) \coloneqq \begin{cases} 1/(N\widetilde{q}_k) & \text{if } N\widetilde{q}_k > 0, \\ 0 & \text{if } N\widetilde{q}_k \leq 0, \end{cases} \qquad S_A(\tau) \coloneqq \big(\big(a_{i_k}^{\mathrm{feat}}, b_{i_k}, \omega_k(\tau)\big)\big)_{k=1}^N .$$

The second case defines S_A on every transcript and leaves the algorithm unchanged, since the sampled weights are positive almost surely. The map S_A is Borel-measurable by its coordinatewise definition. Together with the assumed measurability of $S \mapsto \theta^\star(S)$ in Appendix L.1, this makes $\tau \mapsto \theta^\star(S_A(\tau))$ Borel-measurable. Now let B be obtained by deleting a row of A, with at least one row remaining, and let $f\colon [n-1] \hookrightarrow [n]$ identify each surviving row with its original index. For a transcript $\tau = ((j_k, \widetilde{q}_k))_{k=1}^N$ on B, put $\tau^f \coloneqq ((f(j_k), \widetilde{q}_k))_{k=1}^N$. We then have $S_B(\tau) = S_A(\tau^f)$, and hence $\theta^\star(S_B(\tau)) = \theta^\star(S_A(\tau^f))$. Relabeling the sampled indices by f amounts to extending $p(B)$ by zero at the deleted index. Thus both transcript laws lie on $[n]^N \times \mathbb{R}^N$, and both outputs are obtained through the same Borel-measurable map $\tau \mapsto \theta^\star(S_A(\tau))$.

We now analyze the average sensitivity with the above construction. Fix $\varepsilon \in (0, 1]$ and recall that $p(A)$ is the sampling distribution on $[n]$ defined above. Conditioned on the randomly rounded count taking the value $m \in \mathbb{Z}_{>0}$, the algorithm draws m indices i.i.d. as

$$I_1, \ldots, I_m \sim p(A),$$

and assigns randomly perturbed weights to selected rows. For each $j \in [m]$, conditioned on $I_j = k$, we draw

$$\widetilde{p}_j \sim \mathrm{Unif}([p_k(A), (1 + \varepsilon/2)p_k(A)])$$

independently and set the weight to

$$W_j \coloneqq \frac{1}{m\,\widetilde{p}_j}.$$

If $p_k(A) = 0$, then $\Pr[I_j = k] = 0$ holds and the value of W_j on that event is irrelevant. The full sampling transcript is $((I_j, W_j))_{j=1}^m \in ([n] \times \mathbb{R}_{\geq 0})^m$, and the final ℓ_1-regression step is a deterministic function of this transcript.

Bounding total variation for a single draw (I_j, W_j). Fix $i \in [n]$ and let $A^{(-i)}$ be A with row i removed. First suppose that $p(A^{(-i)})$ is a probability vector, extended by zero at the deleted occurrence. Let P, Q be the one-draw laws of (I_j, W_j) under $A, A^{(-i)}$ at the common count $m \geq 1$. Apply the first bound in Lemma I.6 with $\mu = 1$ and $\beta = \varepsilon/2$. Because $R_1 = 1$ almost surely and $\varepsilon \leq 1$, data processing under the common map $(I, V) \mapsto (I, 1/(mV))$, with weight zero when $V = 0$, gives

$$\mathrm{TV}(P, Q) \leq \frac{5}{2\varepsilon} \|p(A) - p(A^{(-i)})\|_1. \tag{23}$$

From one draw to m draws. Let $P_{A,m}$ and $P_{A^{(-i)},m}$ denote the laws of the count-m transcripts under inputs A and $A^{(-i)}$, respectively. When $p(A^{(-i)})$ is a probability vector, these are the laws of $((I_j, W_j))_{j=1}^m$ and equal $P^{\otimes m}$ and $Q^{\otimes m}$, respectively. Hence Lemma H.2 and (23) give

$$\mathrm{TV}\big(P_{A,m}, P_{A^{(-i)},m}\big) \leq \frac{5m}{2\varepsilon} \|p(A) - p(A^{(-i)})\|_1.$$

If $p(A^{(-i)})$ is the zero vector, the deleted input produces a fixed deterministic output. For the comparison, augment its internal transcript by the same count coordinate and a dummy symbol; this does not change the output and permits a common count-m transcript space. We then have

$$\mathrm{TV}\big(P_{A,m}, P_{A^{(-i)},m}\big) \leq 1 \leq \frac{5m}{2\varepsilon} \|p(A) - p(A^{(-i)})\|_1,$$

because $\|p(A) - p(A^{(-i)})\|_1 = \|p(A)\|_1 = 1$ and $\varepsilon \le 1$. Therefore, for every $m \ge 1$, we have

$$\mathrm{TV}\big(P_{A,m}, P_{A^{(-i)},m}\big) \le \frac{5m}{2\varepsilon} \|p(A) - p(A^{(-i)})\|_1.$$

We now bound the average sensitivity of the randomly rounded ℓ_1 regression algorithm. For each sample count $m \ge 1$, we use the map

$$((I_j, W_j))_{j=1}^m \longmapsto \theta^\star\left(\left(a_{I_j}^{\mathrm{feat}}, b_{I_j}, W_j\right)_{j=1}^m\right)$$

on $([n] \times \mathbb{R}_{\ge 0})^m$. This map is Borel-measurable by the coordinatewise sample construction and the optimizer assumption in Appendix L.1. After relabeling the surviving rows, it gives both outputs; we extend it to send the dummy symbol to $\theta^\star(\emptyset)$. Using data processing, Lemma I.2, and $\mathbb{E}[R_{\mu_\varepsilon}] = \mu_\varepsilon$, we obtain

$$\mathrm{TV}(\mathcal{L}_\varepsilon^{\mathrm{ex}}(A), \mathcal{L}_\varepsilon^{\mathrm{ex}}(A^{(-i)})) \le \frac{5\mu_\varepsilon}{2\varepsilon} \left\| p(A) - p(A^{(-i)}) \right\|_1. \tag{24}$$

Averaging and applying Lemma L.5 gives

$$\frac{1}{n}\sum_{i=1}^n \mathrm{TV}(\mathcal{L}_\varepsilon^{\mathrm{ex}}(A), \mathcal{L}_\varepsilon^{\mathrm{ex}}(A^{(-i)})) \le \frac{5\mu_\varepsilon}{\varepsilon n}.$$

The transcript comparison preceding (24) uses only the two sampling probability vectors and the common output map. We will apply the same comparison to the approximate sampling probabilities in Appendix L.5.

We specify one sensitivity profile for the exact procedure and its approximate-weight implementation. With the mean μ_ε above, choose an absolute constant C_{ℓ_1} satisfying $C_{\ell_1} \ge 140(c_{\ell_1} + 1)$. Then the following function is an average-sensitivity upper bound in Assumption 4:

$$\varphi(\varepsilon) := C_{\ell_1} \frac{d}{\varepsilon^3} \log\left(\mathrm{e} + \frac{dT}{\varepsilon}\right).$$

L.4 Stability under changes in target accuracy

We now keep the input matrix $A \in \mathbb{R}^{n \times d}$ fixed and vary the target accuracy. The sampling distribution $p(A)$ stays fixed; only the sample-count law and the width of the weight perturbation interval change. For $0 < \varepsilon \le \eta \le 1$, we prove

$$\mathrm{TV}\big(\mathcal{L}_\varepsilon^{\mathrm{ex}}(A), \mathcal{L}_\eta^{\mathrm{ex}}(A)\big) = O\left(\mu_\varepsilon \frac{\eta - \varepsilon}{\varepsilon}\right).$$

Together with the profile φ chosen above, this implies Assumption 9 with $C_{\mathrm{pc}} = O(1)$. If A has no non-zero row, the fixed-tie-breaking output is independent of the accuracy parameter, and hence the total variation distance is zero. Assume below that A has at least one non-zero row. Apply Lemma I.7 with $\mathcal{T} = \mathcal{T}^{\mathrm{den}}$, $c = 1/2$, $a = 2$, $C = c_{\ell_1} d$, and $B = 2dT$, using the fixed augmented dimension d in the existing mean μ_ε. The common measurable optimizer and the map $V \mapsto 1/(NV)$ then give

$$\mathrm{TV}\big(\mathcal{L}_\varepsilon^{\mathrm{ex}}(A), \mathcal{L}_\eta^{\mathrm{ex}}(A)\big) \le 4\mu_\varepsilon \frac{\eta - \varepsilon}{\varepsilon}.$$

It remains to convert this same-input bound into Assumption 9. For a dataset X and $x \in X$, let $Y := X \setminus x$ and set $\varepsilon := \varepsilon_X$ and $\eta := \varepsilon_Y$. Since losses are non-negative and bounded by one, we have

$$0 \le \mathrm{OPT}(X) - \mathrm{OPT}(Y) \le 1, \qquad \varepsilon_X \le \varepsilon_Y.$$

Using the above total variation bound on the fixed input Y and the chosen constant C_{ℓ_1}, which makes $\mu_\varepsilon/\varepsilon = O(\varphi(\varepsilon))$, gives

$$\mathrm{TV}\big(\mathcal{L}_{\varepsilon_X}^{\mathrm{ex}}(Y), \mathcal{L}_{\varepsilon_Y}^{\mathrm{ex}}(Y)\big) = O(\varphi(\varepsilon_X)(\varepsilon_Y - \varepsilon_X)).$$

Thus, Lemma J.1, applied with $u = \mathrm{OPT}(X) + \lambda$ and $v = \mathrm{OPT}(Y) + \lambda$, gives

$$\varphi(\varepsilon_X)(\varepsilon_Y - \varepsilon_X) = O(\varepsilon_X^2(\mathrm{OPT}(X) - \mathrm{OPT}(Y))).$$

Since we have $\varepsilon_X \le 1$, this is at most $O(\varepsilon_X(\mathrm{OPT}(X) - \mathrm{OPT}(Y)))$. Averaging over $x \in X$ proves Assumption 9 with $C_{\mathrm{pc}} = O(1)$.

L.5 Polynomial-time implementation with approximate Lewis weights

We now replace exact Lewis weights by computed approximations. For an m-row matrix, we choose the relative accuracy so that normalizing the computed weights changes the sampling distribution by at most $4/m$ in ℓ_1 distance. This preserves the $O(1/n)$ average change under row deletion. The computation is independent of the regression accuracy ε; hence changing ε still affects only the sample count and the weight perturbations. We give the computation first, then verify the approximation, stability and running-time guarantees of the resulting regression procedure.

L.5.1 Computing approximate weights

For an m-row matrix B with non-zero rows, we approximate each Lewis weight within the multiplicative factor

$$\alpha_m := m^{1/(m\log(\mathrm{e}m))} \le \exp(1/m).$$

For $m = 1$, set $\widehat{w}_1(B) := 1$; this equals $w_1(B)$ by Lemma L.2 (2). Suppose henceforth that $m \ge 2$ and let $r := \operatorname{rank}(B)$. To use ordinary inverses in the fixed-point iteration, first reduce B to full column rank: scan its columns in their original order using deterministic Gaussian elimination. Retaining a column precisely when it is linearly independent of the preceding retained columns gives a full-column-rank matrix $C \in \mathbb{R}^{m\times r}$. The same elimination produces matrices $R_B \in \mathbb{R}^{r\times d}$ and $J_B \in \mathbb{R}^{d\times r}$ such that

$$B = CR_B, \qquad C = BJ_B, \qquad R_BJ_B = I_r.$$

Writing $b_j^\top = c_j^\top R_B$ for the corresponding rows of B and C, for every positive diagonal matrix W we have

$$b_j^\top (B^\top W^{-1}B)^\dagger b_j = c_j^\top (C^\top W^{-1}C)^{-1}c_j. \tag{25}$$

Indeed, put $H := C^\top W^{-1}C$ and $y := J_BH^{-1}c_j$. Using $R_BJ_B = I_r$ and $b_j = R_B^\top c_j$, we obtain

$$(B^\top W^{-1}B)y = R_B^\top HR_BJ_BH^{-1}c_j = b_j, \qquad b_j^\top y = c_j^\top H^{-1}c_j.$$

Since $b_j \in \operatorname{range}(B^\top W^{-1}B)$, every solution has the same inner product with b_j, proving (25). Thus compression preserves the Lewis-weight equations and their unique positive solution.

We use the $p = 1$ fixed-point iteration analyzed by Cohen and Peng [17, Lemma 2.4], but perform the following fixed number of approximate square-root updates. For a positive vector v, define

$$F(v)_j := \sqrt{c_j^\top (C^\top \operatorname{diag}(v)^{-1}C)^{-1}c_j}, \qquad j \in [m].$$

Because $B = CR_B$ and B has no zero row, every c_j is non-zero. Thus the matrix inside the inverse is positive definite and every square-root input in the definition of F is strictly positive. Set $v^{(0)} := (1,\dots,1)$ and $\rho_m := 1/(32m^2)$. For $s = 0,\dots,4m-1$, compute each coordinate of $v^{(s+1)}$ with a deterministic square-root oracle so that the following inequalities hold:[9]

$$(1-\rho_m)F(v^{(s)})_j \le v_j^{(s+1)} \le (1+\rho_m)F(v^{(s)})_j \qquad (j \in [m]).$$

We verify that this finite recurrence supplies the required approximation. Let $w := w(C)$ and set $D_s := \left\|\log v^{(s)} - \log w\right\|_\infty$, where the logarithm is coordinatewise. Since C has full column rank and no zero rows, Cohen and Peng [17, Lemma 3.2], applied with $p = 1$, gives, for all positive $u, v \in \mathbb{R}^m$,

$$\|\log F(u) - \log F(v)\|_\infty \le \frac{1}{2}\|\log u - \log v\|_\infty.$$

Moreover, Cohen and Peng [17, Lemma 3.5], applied with $p = 1$ and exact updates ($\beta = 1$ in their notation), gives $\left\|\log F(v^{(0)}) - \log w\right\|_\infty \le \frac{1}{2}\log m$. Since $\rho_m \le 1/2$, one oracle call changes the logarithm of the exact update by at most $\eta_m := 2\rho_m = 1/(16m^2)$. Consequently, we have

$$D_1 \le \frac{1}{2}\log m + \eta_m \le \log m, \qquad D_{s+1} \le \frac{1}{2}D_s + \eta_m. \tag{26}$$

[9] The exact-real model includes a Borel-measurable unit-cost oracle that, given $x > 0$ and a relative tolerance $\rho > 0$, returns $y > 0$ satisfying $(1-\rho)\sqrt{x} \le y \le (1+\rho)\sqrt{x}$.

Iterating (26) and using $2^{-(4m-1)} \le (2m(1+\log m))^{-1}$ and $2\eta_m \le \log m/(2m(1+\log m))$ for $m \ge 2$ gives

$$D_{4m} \le 2^{-(4m-1)} \log m + 2\eta_m \le \frac{\log m}{m(1+\log m)} = \log \alpha_m.$$

Therefore $\widehat{w}(B) \coloneqq v^{(4m)}$, indexed by the rows of B, satisfies

$$\alpha_m^{-1} w_j(B) \le \widehat{w}_j(B) \le \alpha_m w_j(B) \qquad (j \in [m]). \tag{27}$$

The column compression uses $O(mdr)$ exact-real operations and comparisons. Each update forms and inverts an $r \times r$ Gram matrix, evaluates m quadratic forms, and makes m square-root-oracle calls. Thus all $4m$ updates use $O(m^2r^2 + mr^3 + m^2) = O(m^4)$ operations and oracle calls. The resulting $\widehat{w}(B)$, and hence its normalized vector, is deterministic, polynomial-time computable and independent of the later target accuracy ε in the exact-real model.

Normalize the computed weights to obtain the sampling probabilities

$$q_j(B) \coloneqq \frac{\widehat{w}_j(B)}{\sum_k \widehat{w}_k(B)}.$$

For an input whose augmented rows are all zero, define $q(B)$ to be zero on its original row index set. For $m \ge 1$, summing (27) and normalizing gives

$$\alpha_m^{-2} p_j(B) \le q_j(B) \le \alpha_m^2 p_j(B), \qquad \|q(B) - p(B)\|_1 \le \alpha_m^2 - 1 \le \frac{4}{m}. \tag{28}$$

For the last inequality, the case $m = 1$ is exact; for $m \ge 2$, use $\alpha_m^2 \le \exp(2/m)$ and convexity of the exponential function on $[0,1]$ to obtain $\exp(2/m) - 1 \le 4/m$.

L.5.2 Guarantees for regression with approximate weights

Use the offline procedure above with $q(A)$ in place of $p(A)$ in both label sampling and the perturbation intervals, retaining the same mean μ_ε and fixed optimizer. Let $\mathcal{L}_\varepsilon(A)$ denote its output law. This is the polynomial-time procedure used in Section 6. We verify the regression approximation and both stability guarantees for these computed probabilities. For a nonempty retained matrix with $r = \operatorname{rank}(A)$, (28) gives

$$q_i(A) \ge \alpha_n^{-2} \frac{w_i(A)}{r} \ge \mathrm{e}^{-2} \frac{w_i(A)}{r} \qquad (i \in [n]).$$

Apply Lemma L.4 with this probability vector. Conditioning on the rounded count and applying Remark K.2 exactly as for $p(A)$ gives

$$\mathbb{E}_{\theta \sim \mathcal{L}_\varepsilon(A)}[\|A(\theta,1)\|_1] \le (1+\varepsilon) \min_{\theta \in \Theta} \|A(\theta,1)\|_1 + \frac{n}{2T}. \tag{29}$$

Zero augmented rows do not change either objective, and the empty retained input has zero loss. We next control the change in $q(A)$ under row deletion.

For each $i \in [n]$, extend $q(A^{(-i)})$ by zero in the deleted coordinate, just as for $p(A^{(-i)})$.

Lemma L.6. *For $n \ge 1$, we have*

$$\frac{1}{n} \sum_{i=1}^{n} \left\| q(A) - q(A^{(-i)}) \right\|_1 \le \frac{14}{n}.$$

Proof. If $n = 1$, the construction computes the sole Lewis weight exactly, while the deleted matrix satisfies $q(A^{(-i)}) = 0$. Thus the left-hand side is one, which is at most $14/n$. Suppose that $n \ge 2$. By the triangle inequality, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \left\| q(A) - q(A^{(-i)}) \right\|_1 \le \|q(A) - p(A)\|_1 + \frac{1}{n} \sum_{i=1}^{n} \left\| p(A) - p(A^{(-i)}) \right\|_1 \\ + \frac{1}{n} \sum_{i=1}^{n} \left\| p(A^{(-i)}) - q(A^{(-i)}) \right\|_1. \end{aligned}$$

The first term is at most $4/n$ by (28), and the second is at most $2/n$ by Lemma L.5. Each row-deleted matrix has $n-1$ non-zero rows; consequently, the same normalized sandwich bounds the last term by $4/(n-1) \le 8/n$. Summing the three bounds proves the claim. □

Proposition L.7. *Let $A \in \mathbb{R}^{n\times d}$ with $n \geq 1$, fix $\varepsilon \in (0,1]$, and let the randomly rounded algorithm use mean sample count μ_ε. Let $\mathcal{L}_\varepsilon(A)$ be the output law of the resulting offline ℓ_1 regression algorithm. Then, we have*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{TV}\Big(\mathcal{L}_\varepsilon(A), \mathcal{L}_\varepsilon(A^{(-i)})\Big) \leq \frac{35\mu_\varepsilon}{\varepsilon n}.$$

Proof. The transcript comparison used to prove (24) applies to $q(A)$ and its zero-extended row-deleted probability vector as well. It gives

$$\mathrm{TV}(\mathcal{L}_\varepsilon(A), \mathcal{L}_\varepsilon(A^{(-i)})) \leq \frac{5\mu_\varepsilon}{2\varepsilon}\left\|q(A) - q(A^{(-i)})\right\|_1.$$

Averaging and applying Lemma L.6 gives the coefficient $(5/2)\cdot 14 = 35$. Moreover, $\varepsilon \leq 1$ and the definition of μ_ε give $\mu_\varepsilon \leq 2(c_{\ell_1}+1)d\varepsilon^{-2}\log(\mathrm{e}+dT/\varepsilon)$, using $\log(\mathrm{e}+2dT/\varepsilon) \leq 2\log(\mathrm{e}+dT/\varepsilon)$. Thus $C_{\ell_1} \geq 140(c_{\ell_1}+1)$ ensures average sensitivity at most $\varphi(\varepsilon)/(2n)$. The zero-row discard convention extends the same bound to the original input size, completing the proof. □

Proposition L.8. *Let $\mathcal{L}_\varepsilon(A)$ be the output law of the randomly rounded Lewis-weight sampling algorithm on a fixed input matrix $A \in \mathbb{R}^{n\times d}$, with mean sample count $\mu_\varepsilon = c_{\ell_1} d\varepsilon^{-2}\log(\mathrm{e}+2dT/\varepsilon)+2$. For $0 < \varepsilon \leq \eta \leq 1$, it holds that*

$$\mathrm{TV}(\mathcal{L}_\varepsilon(A), \mathcal{L}_\eta(A)) = O\left(\mu_\varepsilon \frac{\eta-\varepsilon}{\varepsilon}\right).$$

Consequently, with $\varphi(\varepsilon) = C_{\ell_1} d\varepsilon^{-3}\log(\mathrm{e}+dT/\varepsilon)$ for a sufficiently large constant $C_{\ell_1} > 0$, the randomly rounded Lewis-weight sampling family satisfies Assumption 9 with $C_{\mathrm{pc}} = O(1)$.

Proof. The computed vector $q(A)$ is independent of ε. The argument in Appendix L.4 therefore applies with $q(A)$ in place of $p(A)$: only the rounded-count mean and the perturbation width change. The optimizer is unchanged, and an input with no non-zero row still returns the same fixed decision. Thus the same-input total-variation bound and its conversion by Lemma J.1 give both claims. □

The approximation bound (29) and the stability bounds in Propositions L.7 and L.8 allow us to apply Corollary 12 to the approximate-weight procedure, yielding the regret bound in Theorem 15. It remains to bound the running time of this procedure.

L.5.3 Running time

At round t, apply the fixed optimizer from Section 6 to the unit-weight prefix and evaluate its objective value to obtain OPT_t, then compute $\varepsilon_t = \varepsilon_{\min}(\mathrm{OPT}_t + \lambda)$. Here $\lambda = H_T^{-4/3}$ is the scalar parameter supplied to Algorithm 1. The optimizer call is polynomial under the exact-optimization assumption in Section 6; evaluating its objective costs $O(td)$, and evaluating $\varepsilon_{\min}$ has unit cost, as assumed in Footnote 5. Let B be the retained m-row augmented matrix at this round, let $r \coloneqq \mathrm{rank}(B)$, put $\varepsilon_\star \coloneqq \varepsilon_{\min}(T+\lambda)$, and define

$$M_{d,T} \coloneqq \lceil \mu_{\varepsilon_\star} \rceil.$$

By Definition 6, $\varepsilon_\star$ minimizes $(T+\lambda)\varepsilon + \varphi(\varepsilon)$ over $0 < \varepsilon \leq 1$, where $\varphi(\varepsilon) = C_{\ell_1} d\varepsilon^{-3}\log(\mathrm{e}+dT/\varepsilon)$ as in Proposition L.8. Comparing the objective value at $\varepsilon_\star$ with that at $\varepsilon = 1$ and using $\log(\mathrm{e}+dT/\varepsilon_\star) \geq 1$ gives

$$C_{\ell_1} d\varepsilon_\star^{-3} \leq T + \lambda + C_{\ell_1} d\log(\mathrm{e}+dT).$$

Hence $\varepsilon_\star^{-1}$, and therefore $M_{d,T}$, is bounded by a polynomial in d and T. When $m > 0$, write N_t for the count N_ε drawn by the offline procedure at round t with $\varepsilon = \varepsilon_t$. Lemma I.4 then gives $N_t \leq M_{d,T}$ for every realization. We treat exact evaluation of the sample-count mean μ_{ε_t} as one unit-cost scalar-oracle call; this convention includes the logarithm in its definition. After that evaluation, draw $U \sim \mathrm{Unif}([0,1))$ and find $N_t = \lfloor \mu_{\varepsilon_t} + U \rfloor$ using at most $M_{d,T}+1$ integer comparisons. The convention in Appendix I assigns unit cost to each draw from an explicitly specified finite or uniform distribution. Thus constructing a cumulative table for $q(B)$, drawing N_t labels, drawing the N_t perturbations from their specified intervals, and forming the finite ordered weighted sample

require $O(m + N_t d + N_t)$ operations and primitive draws. The algorithm then applies the same fixed optimizer to that sample. Together with the $O(mdr + m^4)$ weight computation and $m \le t \le T$, these bounds give a polynomial worst-case cost per round in the exact-real model, uniformly over the realized randomness. If $m = 0$, the algorithm performs the zero-row scan and returns $\theta^\star(\emptyset)$ without sampling or computing Lewis weights; this case also has polynomial cost.

M Details of other applications

We present offline algorithms for online (k, z)-clustering and online low-rank matrix approximation, and establish their approximation, average-sensitivity, and parameter-change stability guarantees.

Below, the assumption that losses lie in $[0, 1]$ is imposed on the instance and the feasible decision class.

Useful common calculus. We will use the following elementary calculus to bound how the sample-count means change with the accuracy parameter. For $a > 0$, $B \ge 0$, $C \ge 1$ and $M(\alpha) = C\alpha^{-a} \log(\mathrm{e} + B/\alpha)$ on $(0, 1]$, the function M is non-increasing and satisfies

$$-M'(s) \le (a+1)\frac{M(s)}{s} \qquad (0 < s \le 1).$$

Consequently, for $0 < \varepsilon \le \eta \le 1$, we have

$$|M(\varepsilon) - M(\eta)| \le (a+1)M(\varepsilon)\frac{\eta - \varepsilon}{\varepsilon}. \tag{30}$$

Indeed, the derivative bound follows by differentiating M, using $\log(\mathrm{e} + B/s) \ge 1$, and bounding the derivative of the logarithmic factor by $1/s$; integrating the derivative from ε to η gives (30).

M.1 Online (k, z)-clustering

We next instantiate the framework for online (k, z)-clustering [18]. The construction uses the coreset sampling tools of Huang and Vishnoi [26] and the average-sensitivity analysis of ordered D^z-sampling due to Yoshida and Ito [63].

M.1.1 Problem setup and computational model

Fix $k \in \mathbb{Z}_{>0}$ and $z \ge 1$. Let $\mathcal{K} \subset \mathbb{R}^d$ be a known nonempty compact center domain and set $\Theta := \mathcal{K}^k$, identifying each tuple with its induced center multiset and allowing repeated centers. For a nonempty finite center set C and a labeled multiset Y, write

$$d(y, C) := \min_{c \in C} \|y - c\|_2, \qquad \mathrm{cost}_z(Y, C) := \sum_{y \in Y} d(y, C)^z, \qquad \mathrm{OPT}(Y) := \min_{C \in \Theta} \mathrm{cost}_z(Y, C).$$

For $C \in \Theta$ and an admissible input point x, define the loss by $\ell(C, x) := d(x, C)^z$. With the cumulative-loss notation of Section 2, we have $\ell(C, Y) = \mathrm{cost}_z(Y, C)$ for every $C \in \Theta$ and every labeled multiset Y of admissible input points. For the coreset analysis only, we also use the unrestricted Euclidean optimum

$$\mathrm{OPT}_{\mathbb{R}^d}(Y) := \min_{C \in (\mathbb{R}^d)^k} \mathrm{cost}_z(Y, C).$$

This finite-data minimum is attained, and we have $\mathrm{OPT}_{\mathbb{R}^d}(Y) \le \mathrm{OPT}(Y)$ because $\Theta \subseteq (\mathbb{R}^d)^k$. The online comparator and every returned decision remain restricted to $\Theta = \mathcal{K}^k$. At round t, the learner plays $C_t \in \Theta$, observes x_t, and incurs the loss $\ell(C_t, x_t)$. We assume that the admissible points and $\mathcal{K}$ are scaled so that $\ell(C, x) \in [0, 1]$ for every $C \in \Theta$ and every admissible input point x. Compactness ensures that the finite weighted clustering objectives over Θ below attain their minima.

Remark M.1 (Computational model)**.** The clustering instantiation assumes access to exact values $\mathrm{OPT}(Y)$ for every observed prefix Y and to a deterministic Borel-measurable exact optimizer over Θ, with fixed tie-breaking, for every weighted coreset objective. For the running-time bounds, we also allow unit-cost exact evaluation of logarithms on positive arguments and positive real powers of nonnegative arguments.

M.1.2 Offline coreset construction

We use D^z-sampling to select preliminary centers, which determine the importance-sampling probabilities for a weighted coreset. We then perturb the sample weights and minimize the resulting weighted loss. The standard D^z-sampling procedure selects its first center uniformly from the input and each subsequent center with probability proportional to the zth power of its distance to the centers selected so far; see Arthur and Vassilvitskii [7] and Yoshida and Ito [63, Algorithm 1] for details. We use independent repetitions so that the union of the selected centers has low cost with high probability.

Specifically, fix a constant $c \geq 2$ and set the number of independent preliminary-center runs to R_{pre}, where

$$R_{\mathrm{pre}} \coloneqq \lceil \log_2(c(T+1)) \rceil, \qquad H \coloneqq kR_{\mathrm{pre}}.$$

On a nonempty labeled multiset Y, run ordered D^z-sampling independently R_{pre} times. Each run records an ordered k-tuple of labeled occurrences, with the following conventions for small inputs and zero distances. If $|Y| < k$, each run first selects every occurrence in a fixed deterministic order and fills the remaining positions by repeating a deterministically selected occurrence of Y. Otherwise, whenever all input points have zero distance to the centers selected so far, each remaining position is sampled independently and uniformly from the labeled occurrences of Y. Thus every run is defined for every nonempty input. Let B be the union of the selected numerical centers after duplicate locations are removed in the order induced by the recorded tuples. Assign each labeled occurrence to its nearest center in B, breaking ties by the order on B. The *cell* of $b \in B$ consists of the occurrences assigned to b. Let $h \coloneqq |B|$. Then we have $h \leq H$, and every center in B has a selected occurrence in its nearest-center cell.

For an occurrence y_i, let $b(i) \in B$ be its assigned center and define

$$d_i \coloneqq d(y_i, B)^z, \qquad D \coloneqq \sum_i d_i, \qquad n_b \coloneqq |\{i : b(i) = b\}|.$$

Define the importance-sampling probabilities by

$$p_i^B \coloneqq \begin{cases} \dfrac{d_i/D + 1/n_{b(i)}}{h+1} & \text{if } D > 0, \\ \dfrac{1}{hn_{b(i)}} & \text{if } D = 0. \end{cases} \tag{31}$$

These probabilities sum to one in both cases.

Choose a constant $\overline{A}_{k,z} \geq 1$ such that one ordered D^z-sampling run satisfies

$$\mathbb{E}_C[\mathrm{cost}_z(Y, C)] \leq \overline{A}_{k,z} \mathrm{OPT}_{\mathbb{R}^d}(Y). \tag{32}$$

By Arthur and Vassilvitskii [7, Theorem 5.1], we may take $\overline{A}_{k,z} = 2^{2z}(\log k + 2) = 2^{O(z)} \log(k+1)$; set $\alpha \coloneqq 2\overline{A}_{k,z}$.

For a finite weighted sequence $W = ((x_j, w_j))_{j=1}^s$, with $s \in \mathbb{Z}_{\geq 0}$, $x_j \in \mathbb{R}^d$ and $w_j \geq 0$ for each $j \in [s]$, define $\mathrm{cost}_W(C) \coloneqq \sum_{j=1}^s w_j d(x_j, C)^z$ for $C \in (\mathbb{R}^d)^k$. For $0 < \xi < 1$, we call W a simultaneous ξ-coreset for Y if

$$(1-\xi)\,\mathrm{cost}_z(Y, C) \leq \mathrm{cost}_W(C) \leq (1+\xi)\,\mathrm{cost}_z(Y, C) \quad \text{for every } C \in (\mathbb{R}^d)^k.$$

For target accuracy $\varepsilon \in (0, 1]$, put

$$\xi \coloneqq \frac{\varepsilon}{10}, \qquad \beta \coloneqq \frac{\varepsilon}{10}, \qquad a \coloneqq 5z + 15, \qquad \delta_T \coloneqq \frac{1}{c(T+1)}.$$

For $\delta \in (0, 1/2)$, choose an integer threshold $s_0(\xi, \delta; \alpha, k, h, z) \geq 1$ so that, whenever $\mathrm{cost}_z(Y, B) \leq \alpha \mathrm{OPT}_{\mathbb{R}^d}(Y)$, every fixed sample count at or above this threshold gives a simultaneous ξ-coreset with probability at least $1 - \delta$, using p^B and inverse-probability weights. The quantitative bound and its proof are given in Lemma M.3 and Appendix M.1.7. We use a smooth sample-count mean to control its change with accuracy, while ensuring that both rounded counts meet the threshold for

every possible $h \le H$. Choose fixed constants $K_{z,\alpha,k,H}, L_{\alpha,k,H} \ge 1$, polynomial in α, k, H up to a $(1+z)^{O(z)}$ factor, such that the function

$$\mu(\varepsilon) \coloneqq K_{z,\alpha,k,H}\varepsilon^{-a}\log\left(\mathrm{e} + \frac{L_{\alpha,k,H}T}{\varepsilon}\right) + 2 \tag{33}$$

satisfies, for every $\varepsilon \in (0,1]$,

$$\lfloor \mu(\varepsilon) \rfloor \ge \max_{1 \le h' \le H} s_0(\varepsilon/10, \delta_T; \alpha, k, h', z) \ge 1. \tag{34}$$

Draw $U \sim \mathrm{Unif}([0,1))$ and set the sample count $N \coloneqq \lfloor \mu(\varepsilon) + U \rfloor$, following the randomized-rounding definition in Appendix I. Conditional on $N = s$, draw labeled occurrences $I_1, \dots, I_s$ independently from p^B, set

$$u_j \coloneqq \frac{1}{s p^B_{I_j}}, \qquad \widetilde{u}_j \sim \mathrm{Unif}([u_j, (1+\beta)u_j]),$$

independently, and return the exact minimizer selected by the fixed tie-breaking rule of

$$C \longmapsto \sum_{j=1}^{s} \widetilde{u}_j \ell(C, y_{I_j}) \quad \text{over } C \in \Theta.$$

Let $S \coloneqq ((y_{I_j}, u_j))_{j=1}^s$ denote the weighted sample before perturbation, and let $\widetilde{S} \coloneqq ((y_{I_j}, \widetilde{u}_j))_{j=1}^s$ denote the same sampled occurrences with perturbed weights. On the empty input, the procedure returns a fixed element of Θ and records an empty transcript.

For every finite labeled input Y and $\varepsilon \in (0,1]$, let $\mathcal{L}_\varepsilon(Y)$ denote the law of the returned clustering.

M.1.3 Approximation guarantee

The preliminary centers must have low cost for the coreset guarantee to apply. Independent repetitions make this condition hold with probability at least $1 - \delta_T$.

Lemma M.2. *For every nonempty finite labeled multiset Y, the union B of the R_{pre} independent preliminary-center tuples satisfies*

$$\Pr[\mathrm{cost}_z(Y,B) > \alpha \mathrm{OPT}_{\mathbb{R}^d}(Y)] \le 2^{-R_{\mathrm{pre}}} \le \delta_T.$$

Proof. Let C denote the centers selected in one run. The expected-cost bound (32) remains valid under our sampling conventions. If the residual potential becomes zero, the selected centers already have zero cost, which additional draws cannot increase. If $|Y| < k$, every input location is selected, so the cost is again zero. When $\mathrm{OPT}_{\mathbb{R}^d}(Y) > 0$, Markov's inequality applied to (32), with $\alpha = 2\overline{A}_{k,z}$, gives

$$\Pr[\mathrm{cost}_z(Y,C) > \alpha \mathrm{OPT}_{\mathbb{R}^d}(Y)] \le \frac{1}{2}. \tag{35}$$

When $\mathrm{OPT}_{\mathbb{R}^d}(Y) = 0$, (32) and nonnegativity imply that the cost is zero almost surely, so (35) also holds. Since B contains the centers from every run, its cost is at most that of each run. Thus its cost can exceed $\alpha \mathrm{OPT}_{\mathbb{R}^d}(Y)$ only if every run's cost does. Independence and (35) give

$$\Pr[\mathrm{cost}_z(Y,B) > \alpha \mathrm{OPT}_{\mathbb{R}^d}(Y)] \le 2^{-R_{\mathrm{pre}}} \le \delta_T,$$

where the last inequality follows from $R_{\mathrm{pre}} = \lceil \log_2(c(T+1)) \rceil$ and $\delta_T = 1/(c(T+1))$. □

The sample must preserve the costs of all feasible center tuples simultaneously. The following coreset guarantee supplies this property whenever the preliminary centers satisfy the factor-α cost bound in Lemma M.2. Simultaneous cost preservation allows us to compare the cost of the coreset minimizer with the optimum on the original input.

Lemma M.3 (Coreset guarantee from a bicriteria approximation)**.** *Let Y be a nonempty finite labeled multiset in $\mathbb{R}^d$, and let $B \subseteq \mathbb{R}^d$ be a nonempty finite center set. Fix a nearest-center tie-breaking rule, assume every cell of B on Y is nonempty, and put $h \coloneqq |B| \ge 1$. Suppose*

$$\mathrm{cost}_z(Y,B) \le \alpha \mathrm{OPT}_{\mathbb{R}^d}(Y), \qquad \alpha \ge 1.$$

For $0 < \xi, \delta < 1/2$ and an integer $s \geq 1$, draw s occurrences independently from (31) and give each draw its inverse-probability weight $1/(sp_i^B)$. There is an integer threshold $s_0(\xi, \delta; \alpha, k, h, z) \geq 1$ satisfying

$$s_0(\xi, \delta; \alpha, k, h, z) \leq (1+z)^{O(z)} \operatorname{poly}(\alpha, k, h) \xi^{-(5z+15)} \log\left(\frac{\operatorname{poly}(\alpha, k, h)}{\xi\delta}\right),$$

such that, for every integer $s \geq s_0(\xi, \delta; \alpha, k, h, z)$, with probability at least $1-\delta$ the weighted sample is a simultaneous ξ-coreset for every $C \in (\mathbb{R}^d)^k$.

We defer the lengthy proof to Appendix M.1.7 to avoid interrupting the flow of the argument.

Each perturbed weight lies between its original value and $1+\beta$ times that value. This controls the effect of perturbation on the returned minimizer; the next proposition also accounts for failure of the preliminary-center and coreset guarantees.

Proposition M.4 (Approximation guarantee for clustering)**.** *For every fixed $\varepsilon \in (0, 1]$ and every input Y of size $t \geq 1$, the offline coreset procedure returns $\widehat{C}$ satisfying*

$$\mathbb{E}\Big[\ell(\widehat{C}, Y)\Big] \leq (1+\varepsilon)\mathrm{OPT}(Y) + 2t\delta_T \leq (1+\varepsilon)\mathrm{OPT}(Y) + \frac{t}{T}.$$

Proof. By Lemma M.2, the bicriteria premise in Lemma M.3 fails with probability at most δ_T. Conditioned on that premise, (34), Lemma I.3, and Lemma M.3 show that the unperturbed weighted sample is a simultaneous ξ-coreset for every unrestricted query, and hence for every $C \in \Theta$, except with probability δ_T. For every query C on the joint success event, we have

$$(1-\xi)\operatorname{cost}_z(Y, C) \leq \operatorname{cost}_S(C) \leq (1+\xi)\operatorname{cost}_z(Y, C).$$

The one-sided perturbation of every sampled weight gives

$$\operatorname{cost}_S(C) \leq \operatorname{cost}_{\widetilde{S}}(C) \leq (1+\beta)\operatorname{cost}_S(C).$$

Comparing the exact minimizer $\widehat{C}$ with a compact-domain optimum $C^\star \in \Theta$ attaining $\mathrm{OPT}(Y)$ therefore yields

$$\ell(\widehat{C}, Y) \leq \frac{(1+\beta)(1+\xi)}{1-\xi}\mathrm{OPT}(Y) \leq (1+\varepsilon)\mathrm{OPT}(Y),$$

where the last inequality follows from $\xi = \beta = \varepsilon/10$. On failure, bounded losses give $\ell(\widehat{C}, Y) \leq t$. The two failure events have total probability at most $2\delta_T$, and we have $2\delta_T \leq 1/T$ because $c \geq 2$ and $T \geq 1$. This proves the expected guarantee. □

M.1.4 Stability under deletion

To analyze stability under deletion, we first bound the change in the law of the ordered D^z-sampling runs that select the preliminary centers. We then fix a common center set B and bound the change in the joint law of the sampled occurrences and their perturbed weights. Together, these bounds control the total variation distance between the complete transcript laws and, by deterministic post-processing, between the returned clustering laws. The next lemma bounds the average sensitivity of one preliminary-center run and of the joint law of all R_{pre} independent runs.

Lemma M.5. *Fix the labeled occurrences of Y as an ambient label space. Let $\perp_{\mathrm{pre}}$ be a distinct symbol for an empty-input preliminary-center run, and define the space of preliminary-center tuples and this empty-input outcome by $\Omega_{\mathrm{pre}}^Y := \{\perp_{\mathrm{pre}}\} \sqcup Y^k$. For every labeled submultiset $Z \subseteq Y$, let $\nu^Y(Z)$ be the pushforward to Ω_{pre}^Y of the ordered labeled k-tuple law for one run on Z under the natural inclusion of occurrence labels, using the Y^k component of the disjoint union, and set $\nu^Y(\emptyset)$ to the point mass at $\perp_{\mathrm{pre}}$. Let $t := |Y| \geq 1$. Then we have*

$$\frac{1}{t}\sum_{i \in Y} \mathrm{TV}\big(\nu^Y(Y), \nu^Y(Y \setminus i)\big) \leq \frac{k}{t}.$$

The joint law $\nu^Y(Y)^{\otimes R_{\mathrm{pre}}}$ of the R_{pre} independent tuples satisfies

$$\frac{1}{t}\sum_{i \in Y} \mathrm{TV}\big(\nu^Y(Y)^{\otimes R_{\mathrm{pre}}}, \nu^Y(Y \setminus i)^{\otimes R_{\mathrm{pre}}}\big) \leq R_{\mathrm{pre}}\frac{k}{t}.$$

Proof. Yoshida and Ito [63, Theorem 2.1] prove an $O(k/t)$ average sensitivity bound for D^z-sampling by comparing next-draw laws at common prefixes. We use their one-step calculation and check the zero-residual and small-input conventions to obtain the stated k/t bound.

Assume $t > k$, so both Y and $Y \setminus i$ have at least k occurrences for every $i \in Y$. Fix $r \in \{0, \dots, k-1\}$ and a prefix $a_{1:r} := (a_1, \dots, a_r) \in Y^r$, where $a_{1:0} := ()$ is the empty tuple. Put $I(a_{1:r}) := \{a_1, \dots, a_r\}$, with $I(a_{1:0}) := \emptyset$. Let $p_{a_{1:r}}$ be the next-draw law on Y after $a_{1:r}$. For each $i \in Y \setminus I(a_{1:r})$, let $q_{a_{1:r},i}$ be the next-draw law on $Y \setminus i$ after the same prefix, embedded in Y with $q_{a_{1:r},i}(i) = 0$. At prefixes of probability zero, use the sampling rule to specify these laws. For a nonempty prefix with positive residual totals on both inputs, the calculation in the proof of Yoshida and Ito [63, Lemma 2.2], with our half-ℓ_1 normalization of total variation, gives

$$\mathrm{TV}(p_{a_{1:r}}, q_{a_{1:r},i}) = p_{a_{1:r}}(i). \tag{36}$$

This calculation uses only the residual weights of the labeled occurrences, so it also applies when locations coincide.

For the initial draw or when the residual total is already zero, both laws are uniform on their respective inputs, and hence

$$\mathrm{TV}(p_{a_{1:r}}, q_{a_{1:r},i}) = \frac{1}{2}\left(\frac{1}{t} + (t-1)\left(\frac{1}{t-1} - \frac{1}{t}\right)\right) = \frac{1}{t} = p_{a_{1:r}}(i).$$

If deletion makes a positive residual total zero, $p_{a_{1:r}}$ is concentrated at i, whereas $q_{a_{1:r},i}(i) = 0$, so $\mathrm{TV}(p_{a_{1:r}}, q_{a_{1:r},i}) = 1 = p_{a_{1:r}}(i)$. Thus (36) holds in all cases. Summing it over deletions outside the prefix gives

$$\sum_{i \in Y \setminus I(a_{1:r})} \mathrm{TV}(p_{a_{1:r}}, q_{a_{1:r},i}) = \sum_{i \in Y \setminus I(a_{1:r})} p_{a_{1:r}}(i) \le 1. \tag{37}$$

For each nonempty labeled submultiset $Z \subseteq Y$ and $r \in \{0, \dots, k\}$, let $\nu_r^Y(Z)$ denote the law on Y^r of the first r draws on Z. In Lemma H.3, the coefficient of the conditional distance at $a_{1:r}$ is $\min\{\nu_r^Y(Y)(a_{1:r}), \nu_r^Y(Y \setminus i)(a_{1:r})\}$. It is zero when $i \in I(a_{1:r})$ and is at most $\nu_r^Y(Y)(a_{1:r})$ otherwise. Applying that lemma, summing over $i \in Y$, and using (37) and $\sum_{a_{1:r} \in Y^r} \nu_r^Y(Y)(a_{1:r}) = 1$ gives, for $0 \le r < k$,

$$\sum_{i \in Y} \mathrm{TV}\big(\nu_{r+1}^Y(Y), \nu_{r+1}^Y(Y \setminus i)\big) \le \sum_{i \in Y} \mathrm{TV}\big(\nu_r^Y(Y), \nu_r^Y(Y \setminus i)\big) + 1. \tag{38}$$

At $r = 0$, all prefix laws are concentrated on the empty tuple, so the sum of their total variation distances is zero. Iterating (38) bounds the sum at $r = k$ by k. Since $\nu_k^Y(Z)$ is $\nu^Y(Z)$ on the Y^k component, division by t proves the claimed k/t bound. For $t \le k$, total variation is at most $1 \le k/t$.

Finally, we bound the deletion sensitivity of the joint law of all R_{pre} runs. For each $i \in Y$, Lemma H.2 gives

$$\mathrm{TV}\big(\nu^Y(Y)^{\otimes R_{\mathrm{pre}}}, \nu^Y(Y \setminus i)^{\otimes R_{\mathrm{pre}}}\big) \le R_{\mathrm{pre}} \mathrm{TV}\big(\nu^Y(Y), \nu^Y(Y \setminus i)\big). \tag{39}$$

Averaging (39) over $i \in Y$ and using the one-run k/t bound proves the claimed $R_{\mathrm{pre}} k/t$ bound. □

We next fix the center set B and its tie-breaking rule and compare the importance-sampling probabilities in (31) before and after one deletion. Provided that the remaining input is nonempty, the next lemma shows that the total variation distance is exactly the original sampling probability of the deleted occurrence.

Lemma M.6 (Deletion identity for fixed-center sampling)**.** *Fix a nonempty finite center set $B \subseteq \mathbb{R}^d$ and its tie-breaking rule. For every nonempty labeled Y, discard the empty cells of B and define p_Y^B by (31), using the resulting number of nonempty cells. For $i \in Y$ with $Y \setminus i \neq \emptyset$, embed $p_{Y \setminus i}^B$ in the labels of Y with mass zero at i. Then we have*

$$\|p_Y^B - p_{Y \setminus i}^B\|_1 = 2p_Y^B(i).$$

Hence $\mathrm{TV}\left(p_Y^B, p_{Y \setminus i}^B\right) = p_Y^B(i)$. *Consequently, for $|Y| \ge 2$, we have*

$$\sum_{i \in Y} \|p_Y^B - p_{Y \setminus i}^B\|_1 = 2.$$

Proof. Delete i and consider a surviving label j. We will confirm $p^B_{Y\setminus i}(j) \geq p^B_Y(j)$. If $D > 0$ and $D - d_i > 0$, then we have $d_j/(D - d_i) \geq d_j/D$; its cell size cannot increase, and the number of nonempty cells cannot increase. Thus its probability cannot decrease. If $D > 0$ and $D - d_i = 0$, all surviving residuals are zero. The residual term disappears while the denominator decreases from one plus the number of cells to at most the number of cells, and hence the same monotonicity holds. If $D = 0$, both residual terms are absent, while cell sizes and the number of nonempty cells again cannot increase. Thus we have $p^B_{Y\setminus i}(j) \geq p^B_Y(j)$ for every surviving label j, and we obtain

$$\|p^B_Y - p^B_{Y\setminus i}\|_1 = p^B_Y(i) + \sum_{j\in Y\setminus i} \Big(p^B_{Y\setminus i}(j) - p^B_Y(j)\Big) = 2p^B_Y(i).$$

Dividing the ℓ_1 identity by two yields the total-variation identity. Summing the ℓ_1 identity over $i \in Y$ gives the final equality, since $\sum_{i\in Y} p^B_Y(i) = 1$. □

We next take the random weight of each sampled occurrence into consideration. For a fixed sample count, the weight assigned to an occurrence is drawn uniformly from an interval whose endpoints are inversely proportional to its sampling probability. The next lemma bounds the total variation distance between the joint laws of the sampled occurrence and its weight in terms of the ℓ_1 distance between the sampling probability vectors.

Lemma M.7 (Total variation bound for sampling with perturbed weights)**.** *Let p, q be probability vectors on a common finite index set, let $b > 0$, and let $0 < \beta \leq 1$. For each $r \in \{p, q\}$, define $\mathcal{J}_{\beta,r}$ as the joint law of (I, W) obtained by drawing $I \sim r$ and then sampling W conditionally on I according to*

$$W \mid I = i \sim \mathrm{Unif}\left(\left[\frac{b}{r_i}, \frac{(1+\beta)b}{r_i}\right]\right) \qquad \text{for } r_i > 0.$$

For $r_i = 0$, take the conditional law of W to be the point mass at zero; these indices are sampled with probability zero. Then we have

$$\mathrm{TV}(\mathcal{J}_{\beta,p}, \mathcal{J}_{\beta,q}) \leq \frac{2+3\beta}{2\beta}\|p - q\|_1.$$

Proof. For direct weight perturbation with count one, the proof of Lemma I.6 gives the bound $(1/2 + (1+\beta)/\beta)\|p - q\|_1$. Scaling the weight by the common positive factor b gives exactly $\mathcal{J}_{\beta,p}$ and $\mathcal{J}_{\beta,q}$. Data processing and $1/2 + (1+\beta)/\beta = (2+3\beta)/(2\beta)$ give the claimed bound. □

Lemma M.5 bounds the average change in the law of the preliminary-center tuples under deletion. For fixed centers, Lemmas M.6 and M.7 bound the change in the joint law of a sampled occurrence and its weight. The next proposition combines these comparisons over all draws to bound the average sensitivity of the complete transcript. Deterministic post-processing gives the same bound for the returned clustering.

Fix $\varepsilon \in (0, 1]$ and a nonempty finite labeled multiset Y. To compare executions on labeled submultisets $Z \subseteq Y$, use their original occurrence labels from Y: deletion removes the deleted occurrence's label, while every surviving label still refers to its original data point. For inputs of size less than k, the repeated occurrences completing each preliminary tuple are chosen using the inherited order on Y. Nearest-center ties are resolved by the order on B specified in the construction.

These labels place all transcripts in a common space. Let $\perp_{\mathrm{tr}}$ be a distinct symbol for the complete sampling transcript on empty input, and let $\Omega_Y := \{\perp_{\mathrm{tr}}\} \sqcup ((\{\perp_{\mathrm{pre}}\} \sqcup Y^k)^{R_{\mathrm{pre}}} \times \bigsqcup_{s\in\mathbb{Z}_{\geq 0}} (Y \times \mathbb{R})^s)$ be equipped with its product and disjoint-union measurable structure. For every nonempty labeled submultiset $Z \subseteq Y$, let $\mathcal{T}^Y_\varepsilon(Z)$ be the law of the transcript generated on Z, expressed using these labels in the product component of Ω_Y; set $\mathcal{T}^Y_\varepsilon(\emptyset)$ to the point mass at $\perp_{\mathrm{tr}}$.

Proposition M.8 (Average sensitivity of the clustering sampling transcript)**.** *For the transcript laws defined above and $t = |Y| \geq 1$, we have*

$$\frac{1}{t}\sum_{i\in Y} \mathrm{TV}\big(\mathcal{T}^Y_\varepsilon(Y), \mathcal{T}^Y_\varepsilon(Y \setminus i)\big) \leq \frac{H}{t} + \frac{23\mu(\varepsilon)}{\varepsilon t}.$$

The same bound holds for the output laws of the clustering procedure.

Proof. The case $t = 1$ follows from the trivial bound one. We may therefore assume that $t \geq 2$. Fix $i \in Y$. Concatenating the R_{pre} preliminary k-tuples in run order gives a sequence of length $H = kR_{\mathrm{pre}}$. Write $a_{1:H} \coloneqq (a_1, \ldots, a_H) \in Y^H$ for this sequence, and let P and Q_i be its laws on inputs Y and $Y \setminus i$, respectively. Apply Lemma H.3 to the two transcript laws, taking this sequence as the first coordinate. The conditional distance at $a_{1:H}$ then has coefficient $\min\{P(a_{1:H}), Q_i(a_{1:H})\}$. If $a_{1:H}$ contains i, then $Q_i(a_{1:H}) = 0$, so this coefficient is zero. For each positive coefficient, the same $a_{1:H}$ therefore determines the same B and its tie-breaking order on both inputs. Each center retains a selected occurrence in its cell, so every cell remains nonempty after deletion. The rounded-count laws are identical because ε is fixed.

Conditional on a common count s, one sampled occurrence–weight pair has the law in Lemma M.7 with $b = 1/s$. Since $\beta = \varepsilon/10$ and $\varepsilon \leq 1$, that lemma and Lemma H.2 bound the total variation distance between the two conditional laws of the s occurrence–weight pairs by

$$\frac{23s}{2\varepsilon} \|p_Y^B - p_{Y\setminus i}^B\|_1.$$

For the conditional contribution, we use $\min\{P(a_{1:H}), Q_i(a_{1:H})\} \leq P(a_{1:H})$. For each $a_{1:H}$, Lemma M.6 bounds the sum over deletions. Averaging over $a_{1:H} \sim P$ and the rounded count, and dividing by t, gives a contribution at most

$$\frac{23\mathbb{E}[N]}{\varepsilon t} = \frac{23\mu(\varepsilon)}{\varepsilon t}.$$

The preliminary contribution is at most $R_{\mathrm{pre}}k/t = H/t$ by Lemma M.5. Adding the two contributions proves the transcript bound.

For all nonempty $Z \subseteq Y$, the same coordinatewise map replaces each sampled label by its fixed data point to form the ordered weighted sample. This map is measurable, and the exact optimizer is Borel-measurable (as assumed in Remark M.1). With the prescribed output on $\perp_{\mathrm{tr}}$ and fixed values on unused components, their composition is a measurable function on Ω_Y. Data processing therefore gives the same bound for the output laws. □

To express the deletion bound in Proposition M.8 as a sensitivity profile, put

$$q \coloneqq 5z + 16, \qquad C_1 \coloneqq 2H + 46(K_{z,\alpha,k,H} + 2),$$

and define the sensitivity profile

$$\overline{\varphi}_{\mathrm{cl}}(\varepsilon) \coloneqq C_1 \varepsilon^{-q} \log\left(\mathrm{e} + \frac{L_{\alpha,k,H} T}{\varepsilon}\right). \tag{40}$$

For $\varepsilon \in (0, 1]$, we have $\mu(\varepsilon) \leq (K_{z,\alpha,k,H} + 2)\varepsilon^{-(q-1)} \log(\mathrm{e} + L_{\alpha,k,H}T/\varepsilon)$. Thus, for every $t \geq 1$, we have

$$\frac{H}{t} + \frac{23\mu(\varepsilon)}{\varepsilon t} \leq \frac{\overline{\varphi}_{\mathrm{cl}}(\varepsilon)}{2t}. \tag{41}$$

Because $H = kR_{\mathrm{pre}} = O(k \log T)$ and $K_{z,\alpha,k,H}$ is polynomial in H, the horizon dependence in C_1 is polylogarithmic.

Proposition M.9 (Average sensitivity of the returned clustering)**.** *For every fixed $\varepsilon \in (0, 1]$ and every input Y of size $t \geq 1$, the output law $\mathcal{L}_\varepsilon(Y)$ satisfies*

$$\frac{1}{t} \sum_{i \in Y} \mathrm{TV}(\mathcal{L}_\varepsilon(Y), \mathcal{L}_\varepsilon(Y \setminus i)) \leq \frac{\overline{\varphi}_{\mathrm{cl}}(\varepsilon)}{2t}.$$

Proof. The output-law bound in Proposition M.8, together with (41), gives the claimed average-sensitivity bound. □

M.1.5 Stability under changes in target accuracy

For a fixed input, the preliminary-center law and the row sampling probabilities do not depend on the target accuracy. We therefore need to compare only the rounded counts and the weight perturbations.

Proposition M.10 (Parameter-change stability for clustering)**.** *For every fixed input Y and $0 < \varepsilon \le \eta \le 1$, we have*

$$\mathrm{TV}(\mathcal{L}_\varepsilon(Y), \mathcal{L}_\eta(Y)) \le (5z+17)\mu(\varepsilon)\frac{\eta-\varepsilon}{\varepsilon} = O_z(\overline{\varphi}_{\mathrm{cl}}(\varepsilon)(\eta-\varepsilon)).$$

Consequently, Assumption 9 holds with $C_{\mathrm{pc}} = O_z(1)$.

Proof. If $Y = \emptyset$, both laws are the same point mass. We may therefore assume that Y is nonempty. The preliminary tuples and the sampling law p^B are independent of the target accuracy, and we therefore couple them identically. Applying Lemma I.7 conditionally on the common preliminary tuples, with $\mathcal{T} = \mathcal{T}^{\mathrm{wt}}$, $c = 1/10$, and $a = 5z+15$, gives

$$\mathrm{TV}(\mathcal{L}_\varepsilon(Y), \mathcal{L}_\eta(Y)) \le (5z+17)\mu(\varepsilon)\frac{\eta-\varepsilon}{\varepsilon}.$$

Here the mean in (33) uses the fixed bound H, independently of the realized number of preliminary centers. Averaging under the common accuracy-independent preliminary law and applying the common measurable output map preserves the bound. The second bound follows from (40).

For the adaptive statement, use the rule (1) with $\overline{\varphi}_{\mathrm{cl}}$, fix X, delete x, write $Y = X\setminus x$, and put $\varepsilon := \varepsilon_X$ and $\eta := \varepsilon_Y$, with $\varepsilon \le \eta$. Nonnegative losses bounded by one give $0 \le \mathrm{OPT}(X) - \mathrm{OPT}(Y) \le 1$. Applying Lemma J.1 to $\overline{\varphi}_{\mathrm{cl}}$ yields

$$\begin{aligned}\overline{\varphi}_{\mathrm{cl}}(\varepsilon_X)(\varepsilon_Y - \varepsilon_X) &= \varepsilon_X \overline{\varphi}_{\mathrm{cl}}(\varepsilon_X)\frac{\varepsilon_Y - \varepsilon_X}{\varepsilon_X} \\ &= O_z\big(\varepsilon_X^2(\mathrm{OPT}(X) - \mathrm{OPT}(Y))\big) \\ &= O_z(\varepsilon_X(\mathrm{OPT}(X) - \mathrm{OPT}(Y))).\end{aligned}$$

Applying the parameter-change bound proved above with $(\varepsilon, \eta) = (\varepsilon_X, \varepsilon_Y)$ and then using this estimate gives

$$\mathrm{TV}(\mathcal{L}_{\varepsilon_X}(Y), \mathcal{L}_{\varepsilon_Y}(Y)) = O_z(\varepsilon_X(\mathrm{OPT}(X) - \mathrm{OPT}(Y))).$$

Averaging over $x \in X$ proves Assumption 9. □

M.1.6 Regret and computational complexity

For fixed z, Propositions M.4, M.9 and M.10 verify the hypotheses of Theorem 11 with $\overline{\varphi}_{\mathrm{cl}}$ defined in (40). In particular, we have

$$\overline{\varphi}_{\mathrm{cl}}(\varepsilon) = \widetilde{O}_{k,z}\Big(\varepsilon^{-(5z+16)}\log(\mathrm{e}/\varepsilon)\Big).$$

Apply Corollary 12 with

$$q = 5z+16, \qquad \lambda = H_T^{-(q+1)/q}, \qquad C_2 = L_{\alpha,k,H}T.$$

This gives

$$\mathcal{R}_T = \widetilde{O}\Big(C_1 + C_1^{1/(5z+17)}\Big(1 + \mathrm{OPT}_T^{(5z+16)/(5z+17)}\Big)\Big) = \widetilde{O}_{k,z}\Big(1 + \mathrm{OPT}_T^{(5z+16)/(5z+17)}\Big).$$

Let N_t denote the number of occurrences in the weighted sample at prefix t, and put $\varepsilon_\star := \varepsilon_{\min}(T+\lambda)$. For every online history, Lemma I.4 gives the deterministic bound

$$N_t \le \lceil \mu(\varepsilon_\star) \rceil \quad \text{almost surely.}$$

Comparing the objective defining $\varepsilon_\star$ with the feasible point $\varepsilon = 1$ and using (40) gives $\varepsilon_\star^{-(5z+16)} \le (T+\lambda)/C_1 + \log(\mathrm{e} + L_{\alpha,k,H}T)$; hence, for fixed z, the bound on N_t is polynomial in k and T.

Under the computational model of Remark M.1, each prefix t requires one oracle call to evaluate $\mathrm{OPT}(X_t)$ exactly and one oracle call to minimize the weighted coreset objective exactly. The preliminary D^z-sampling consists of exactly R_{pre} runs. Forming the preliminary tuples, cells, sampling probabilities, and weighted sample takes polynomially many exact-real operations in $t, d, k, R_{\mathrm{pre}}$ and the bound on N_t.

M.1.7 Proof of the coreset guarantee

Recall the setting of Lemma M.3: B is a fixed set of $h := |B|$ centers with nonempty nearest-center cells on the labeled input Y. For each occurrence $i \in Y$, $b(i)$ is its assigned center, and n_b is the size of the cell of $b \in B$. The residual costs satisfy

$$d_i := d(y_i, B)^z, \qquad D := \sum_{i \in Y} d_i = \mathrm{cost}_z(Y, B) \le \alpha \mathrm{OPT}_{\mathbb{R}^d}(Y).$$

Here $\alpha \ge 1$, and $\mathrm{OPT}_{\mathbb{R}^d}(Y)$ is the minimum cost over all k-center tuples in $\mathbb{R}^d$. For a fixed integer $s \ge 1$, the weighted sample S consists of s independent draws from p^B in (31), each draw of occurrence i receiving weight $1/(sp_i^B)$. The parameters $0 < \xi, \delta < 1/2$ specify the coreset accuracy and allowed failure probability, respectively.

The proof reduces approximation of all center costs to two guarantees: cost preservation in a fixed low-dimensional subspace and a bound on the error of replacing an arbitrary center tuple by a representative in that subspace. After treating $D = 0$, we deduce the coreset conclusion from these guarantees (Part I). We then construct the subspace and representatives (Part II), prove cost preservation in it (Part III), and use residual-fitting and projection-cost bounds (Part IV) to control the replacement error (Part V). Finally, we bound the required sample count (Part VI).

Proof of Lemma M.3. First suppose $D = 0$. Every point in a cell is numerically equal to its assigned center, and every cell has probability $1/h$ under (31). Let M_b be the number of draws from cell b, and set $c_0 := 8$. For the case $D = 0$, it suffices to choose the sample-size threshold s_0 so that

$$s_0 \ge c_0 h \xi^{-2} \log\left(\frac{h}{\delta}\right).$$

For every fixed integer $s \ge s_0$, a Chernoff bound and a union bound then give $|M_b - s/h| \le \xi s/h$ simultaneously for all cells with probability at least $1 - \delta$. The total sampled weight in cell b then lies in $(1 \pm \xi) n_b$. Both the original and sampled costs of any query are nonnegative sums of the same values $d(b, C)^z$ with these respective coefficients, proving the simultaneous coreset claim. Below we consider $D > 0$.

Part I: Proof from two cost bounds. Set $\eta := \xi/100$. For each $W \in \{Y, S\}$, write cost_W for the weighted clustering cost, assigning unit weights to the occurrences of Y. We shall construct a subspace $\Gamma' \subseteq \mathbb{R}^d$ containing B and a representative $C' \in (\Gamma')^k$ for each $C \in (\mathbb{R}^d)^k$. The subspace and the map $C \mapsto C'$ depend only on the input and fixed parameters and are chosen before drawing S; their construction is given in Part II. Assume for now that the following two comparisons hold. The first is cost preservation for all center tuples in Γ':

$$(1 - \eta)\,\mathrm{cost}_z(Y, Q) \le \mathrm{cost}_S(Q) \le (1 + \eta)\,\mathrm{cost}_z(Y, Q) \quad \text{for every } Q \in (\Gamma')^k. \tag{42}$$

The second bounds the error of replacing any center tuple $C \in (\mathbb{R}^d)^k$ by its representative. For the same C' and every $W \in \{Y, S\}$, we require

$$\left|\mathrm{cost}_W(C)^{1/z} - \mathrm{cost}_W(C')^{1/z}\right| \le \frac{\eta}{10z} \mathrm{OPT}_{\mathbb{R}^d}(Y)^{1/z}. \tag{43}$$

We prove (42) in Part III and (43) in Part V. The probability that both hold, and the required sample count, are established in Part VI.

For an arbitrary $C \in (\mathbb{R}^d)^k$, set $a := \mathrm{cost}_z(Y, C)^{1/z}$, $a' := \mathrm{cost}_z(Y, C')^{1/z}$, $b := \mathrm{cost}_S(C)^{1/z}$, and $b' := \mathrm{cost}_S(C')^{1/z}$. Using $\mathrm{OPT}_{\mathbb{R}^d}(Y)^{1/z} \le a$, (43) controls both $|a - a'|$ and $|b - b'|$ by $\eta a/(10z)$. The low-dimensional event (42) applies to every $C' \in (\Gamma')^k$ and gives $(1 - \eta)^{1/z} a' \le b' \le (1+\eta)^{1/z} a'$. For $0 < \eta \le 1/200$ and $z \ge 1$, concavity of the z-th root and Bernoulli's inequality give $|b' - a'| \le (2\eta/z) a'$. Together with $a' \le (1 + \eta/(10z))a$, the triangle inequality now gives

$$|b - a| \le \frac{2\eta}{10z} a + \frac{2\eta}{z}\left(1 + \frac{\eta}{10z}\right) a \le \frac{3\eta}{z} a.$$

Bernoulli's inequality gives $(1 - 3\eta/z)^z \ge 1 - 3\eta$. Applying it also to the reciprocal base gives $(1 + 3\eta/z)^z \le (1 - 3\eta)^{-1} \le 1 + 4\eta$. Raising the resulting nonnegative root bounds to the power z therefore yields

$$(1 - 3\eta)\,\mathrm{cost}_z(Y, C) \le \mathrm{cost}_S(C) \le (1 + 4\eta)\,\mathrm{cost}_z(Y, C).$$

The inequality $4\eta \le \xi$ establishes the simultaneous unrestricted coreset claim. Restriction to $C \in \Theta$ is immediate.

Part II: Construction of the comparison subspace. We construct Γ' by first finding a subspace Γ for which adding any k centers can reduce the projection cost only slightly. Set

$$\rho := \frac{\eta^{z+3}}{6(84z)^{2z}}, \qquad \tau := \frac{\rho}{\alpha}. \tag{44}$$

For a linear subspace $\Lambda \subseteq \mathbb{R}^d$, let π_Λ denote orthogonal projection onto Λ and define $d(y,\Lambda) := \|y - \pi_\Lambda y\|_2$ for $y \in \mathbb{R}^d$. For a finite multiset W with nonnegative occurrence weights w_y, define $P_\Lambda(W) := \sum_{y\in W} w_y d(y,\Lambda)^z$, treating Y as unit-weighted. For $C \in (\mathbb{R}^d)^k$, the reduction in this cost when C is adjoined to Λ is

$$\Delta_\Lambda(W,C) := P_\Lambda(W) - P_{\mathrm{span}(\Lambda\cup C)}(W).$$

Start with $\Gamma_0 := \mathrm{span}(B)$ and, while some query C satisfies

$$\Delta_\Gamma(Y,C) > \frac{\tau \mathrm{OPT}_{\mathbb{R}^d}(Y)}{2},$$

replace Γ by $\mathrm{span}(\Gamma \cup C)$. Since

$$P_{\Gamma_0}(Y) \le \mathrm{cost}_z(Y,B) \le \alpha \mathrm{OPT}_{\mathbb{R}^d}(Y),$$

the process performs at most $2\alpha/\tau$ updates. Its terminal subspace satisfies

$$\dim(\Gamma) \le h + \frac{2\alpha k}{\tau}, \qquad \Delta_\Gamma(Y,C) \le \frac{\tau \mathrm{OPT}_{\mathbb{R}^d}(Y)}{2} \quad \text{for every } C \in (\mathbb{R}^d)^k. \tag{45}$$

If $\Gamma \neq \mathbb{R}^d$, adjoin one fixed unit vector $v \in \Gamma^\perp$ and call the resulting space Γ'; otherwise set $\Gamma' := \Gamma$. Thus we have

$$\dim(\Gamma') \le h + \frac{2\alpha k}{\tau} + 1.$$

For each center c, define its representative by

$$c' := \pi_\Gamma c + d(c,\Gamma) v$$

when $\Gamma \neq \mathbb{R}^d$, where v is the fixed unit vector used to define Γ', and set $c' := c$ when $\Gamma = \mathbb{R}^d$. Applying this construction coordinatewise gives $C' \in (\Gamma')^k$. Each c' has the same projection onto Γ and the same distance from Γ as c. The Pythagorean theorem therefore gives

$$d(x,C') = d(x,C) \qquad \text{for every } x \in \Gamma. \tag{46}$$

Part III: Cost preservation in the subspace. For every query $C \in (\mathbb{R}^d)^k$ with $\mathrm{cost}_z(Y,C) > 0$, the relaxed triangle-inequality calculation in Huang and Vishnoi [26, Claim 5.6], applied to the h cells of B, gives

$$\frac{d(y_i,C)^z}{\mathrm{cost}_z(Y,C)} \le 2^z \frac{d_i}{\mathrm{OPT}_{\mathbb{R}^d}(Y)} + \frac{2^{2z+1}\alpha}{n_{b(i)}}.$$

Because $D \le \alpha \mathrm{OPT}_{\mathbb{R}^d}(Y)$ and $2^z \le 2^{2z+1}$, the right-hand side is at most

$$\sigma_B(i) := 2^{2z+1}\alpha\left(\frac{d_i}{D} + \frac{1}{n_{b(i)}}\right).$$

Its normalization is exactly (31). The sum of the sensitivity bounds is

$$\Sigma_B := \sum_{i\in Y} \sigma_B(i) = 2^{2z+1}\alpha(h+1).$$

Apply the sensitivity-sampling argument in Huang and Vishnoi [26, proof of Lemma 5.5, based on Theorem 3.1(2)], with center tuples restricted to $(\Gamma')^k$ and accuracy η. For $y \in \mathbb{R}^d$ and $Q \in (\Gamma')^k$, orthogonal decomposition gives

$$d(y,Q)^2 = d(\pi_{\Gamma'} y, Q)^2 + d(y,\Gamma')^2. \tag{47}$$

By (47), the distance functions for these queries depend only on coordinates along Γ' and the fixed residual distance of each input point. The restricted-query sampling argument therefore uses $\dim(\Gamma')$, up to an absolute constant, in place of the ambient dimension. The bound $\sigma_B(i)$ already dominates the sensitivity of every occurrence for all queries in $(\mathbb{R}^d)^k$, and hence also for this restricted family. Normalizing these bounds gives the sampling probabilities p_i^B. This proves (42) with probability at least $1 - \delta/3$ for every fixed integer $s \ge N_{\mathrm{low}}$, where the sufficient threshold N_{low} is specified in Part VI.

Part IV: Residual fitting and projection cost. We prove a lower bound on the minimum cost of fitting the sampled residuals by a linear subspace of dimension at most k, and an upper bound on the sample's projection cost onto Γ. For a sufficiently large sample count, each bound holds with probability at least $1-\delta/3$. We use these bounds in Part V to control the error of replacing arbitrary centers by their representatives. Put $r_i := y_i - \pi_\Gamma y_i \in \Gamma^\perp$ for each labeled occurrence i of Y. Minimizing over linear subspaces $V \subseteq \Gamma^\perp$ of dimension at most k, define

$$\Psi_Y := \min_{\dim(V)\le k} \sum_i d(r_i, V)^z, \qquad \Psi_S := \min_{\dim(V)\le k} \sum_{j=1}^{s} \frac{d(r_{I_j}, V)^z}{s p^B_{I_j}}.$$

For residual subspace fitting, put $G_{\mathrm{res}} := 2\alpha(h+1)/\tau$ and $\varepsilon_{\mathrm{res}} := \tau/(2\alpha)$. The required lower bound is

$$\Psi_S \ge \Psi_Y - \frac{\tau \mathrm{OPT}_{\mathbb{R}^d}(Y)}{2}. \tag{48}$$

If $\Psi_Y \le \tau\mathrm{OPT}_{\mathbb{R}^d}(Y)/2$, this inequality holds deterministically by nonnegativity. Otherwise, since $B \subseteq \Gamma$ and $D \le \alpha\mathrm{OPT}_{\mathbb{R}^d}(Y)$, for every such subspace V and every labeled occurrence i we have

$$\frac{d(r_i, V)^z}{\sum_j d(r_j, V)^z} \le \frac{d_i}{\Psi_Y} \le G_{\mathrm{res}} p^B_i.$$

Thus the values $G_{\mathrm{res}} p^B_i$ satisfy the sensitivity condition in Huang and Vishnoi [26, Theorem 5.10]; their sum is G_{res}, and their normalization gives exactly p^B. Apply that theorem with accuracy $\varepsilon_{\mathrm{res}}$ and failure probability $\delta/3$. Testing the zero subspace gives $\Psi_Y \le P_\Gamma(Y) \le D \le \alpha\mathrm{OPT}_{\mathbb{R}^d}(Y)$. The multiplicative lower bound therefore implies (48) with probability at least $1-\delta/3$. For the sampled projection cost, Hoeffding's inequality gives

$$P_\Gamma(S) \le P_\Gamma(Y) + \frac{\tau \mathrm{OPT}_{\mathbb{R}^d}(Y)}{2}. \tag{49}$$

Indeed, because $B \subseteq \Gamma$, we have

$$\frac{d(y_i, \Gamma)^z}{p^B_i} \le (h+1)D \le \alpha(h+1)\mathrm{OPT}_{\mathbb{R}^d}(Y).$$

The sufficient sample-count thresholds N_{res} and N_{proj} for these two events are specified in Part VI.

Part V: Comparison with representative centers. Assume the residual-fitting and projection-cost events (48) and (49). We prove (43) by bounding how far the projected points move when a center tuple $C \in (\mathbb{R}^d)^k$ is adjoined to Γ. For every $W \in \{Y, S\}$, define

$$M_W(C) := \sum_{y \in W} w_y \left\| \pi_\Gamma y - \pi_{\mathrm{span}(\Gamma \cup C)} y \right\|_2^z.$$

The residual-fitting and projection-cost events, together with (45), give

$$P_\Gamma(S) \le (\alpha+1)\mathrm{OPT}_{\mathbb{R}^d}(Y), \qquad \Delta_\Gamma(S, C) \le \frac{3\tau\mathrm{OPT}_{\mathbb{R}^d}(Y)}{2} \le 2\tau\mathrm{OPT}_{\mathbb{R}^d}(Y). \tag{50}$$

For $z \ge 2$, the Pythagorean theorem for the nested subspaces $\Gamma \subseteq \mathrm{span}(\Gamma \cup C)$ gives

$$M_W(C) \le \Delta_\Gamma(W, C). \tag{51}$$

For $1 \le z \le 2$, the weighted Cauchy inequality gives

$$M_W(C) \le 2^{z/2} \sqrt{P_\Gamma(W) \Delta_\Gamma(W, C)}. \tag{52}$$

Since $B \subseteq \Gamma$, we have $P_\Gamma(Y) \le D \le \alpha\mathrm{OPT}_{\mathbb{R}^d}(Y)$. Together with (45) and (50), this gives $P_\Gamma(W) \le 2\alpha\mathrm{OPT}_{\mathbb{R}^d}(Y)$ and $\Delta_\Gamma(W, C) \le 2\tau\mathrm{OPT}_{\mathbb{R}^d}(Y)$ for every $W \in \{Y, S\}$, where we use $\alpha \ge 1$. Substituting these bounds and $\tau = \rho/\alpha$ into (51) and (52) gives

$$M_W(C) \le \omega_z \mathrm{OPT}_{\mathbb{R}^d}(Y), \qquad \omega_z := \begin{cases} 4\sqrt{\rho} & \text{if } 1 \le z \le 2, \\ 2\rho & \text{if } z > 2, \end{cases} \qquad 4\omega_z^{1/z} \le \frac{\eta}{10z}. \tag{53}$$

For $1 \le z \le 2$, the current parameters satisfy $\rho \le (\eta/(84z))^{2z}/4$, because we have $\eta \le 1$ and $z + 3 \ge 2z$. Thus, we have $4(4\sqrt{\rho})^{1/z} \le 8\eta/(84z) \le \eta/(10z)$. For $z > 2$, we have $2\rho \le$

$(\eta/(84z))^z$, which gives $4(2\rho)^{1/z} \le 4\eta/(84z) \le \eta/(10z)$. Applying the Pythagorean theorem to the projections onto Γ and $\operatorname{span}(\Gamma \cup C)$, together with the triangle inequality, gives, for every $y \in \mathbb{R}^d$,

$$\left| d(y,C) - \sqrt{d(\pi_\Gamma y, C)^2 + d(y,\Gamma)^2} \right| \le 2 \left\| \pi_{\operatorname{span}(\Gamma \cup C)} y - \pi_\Gamma y \right\|_2. \tag{54}$$

By (46), the square-root term in (54) is the same for C and C'. Thus, for every $W \in \{Y, S\}$, we have

$$\begin{aligned} \left| \operatorname{cost}_W(C)^{1/z} - \operatorname{cost}_W(C')^{1/z} \right| &\le \left(\sum_{y \in W} w_y \left| d(y,C) - d(y,C') \right|^z \right)^{1/z} \\ &\le 2M_W(C)^{1/z} + 2M_W(C')^{1/z}. \end{aligned} \tag{55}$$

The first inequality follows from the reverse triangle inequality applied to $y \mapsto d(y,C)$ and $y \mapsto d(y,C')$ in the weighted ℓ_z norm with weights w_y. The second uses (54) for both tuples, followed by Minkowski's inequality for the same norm. Applying (53) to both terms in (55) proves (43) for every $C \in (\mathbb{R}^d)^k$ and every $W \in \{Y, S\}$.

Part VI: Sample-size bound. We now determine a sample-size threshold that ensures the coreset guarantee with probability at least $1 - \delta$. Combining the requirements for $D = 0$ and $D > 0$, we obtain a common threshold s_0 and bound it as stated in Lemma M.3. For the projection-cost event (49), it is enough to use the sample-count threshold

$$N_{\mathrm{proj}} := \left\lceil 2\alpha^2 (h+1)^2 \tau^{-2} \log \frac{3}{\delta} \right\rceil.$$

Using $\dim(\Gamma') \le h + 2\alpha k/\tau + 1$, we can choose sample-count thresholds N_{low} for the low-dimensional coreset event and N_{res} for residual subspace fitting independently of the input. There are absolute constants $c_{\mathrm{low}}, c_{\mathrm{res}} > 0$ for which we may take

$$\begin{aligned} N_{\mathrm{low}} &:= \left\lceil c_{\mathrm{low}} \eta^{-2} \Sigma_B \left(\left(h + \frac{2\alpha k}{\tau} + 1 \right) k \log \Sigma_B + \log \frac{3}{\delta} \right) \right\rceil, \\ N_{\mathrm{res}} &:= \left\lceil c_{\mathrm{res}} (1+z)^2 G_{\mathrm{res}}^2 \varepsilon_{\mathrm{res}}^{-2} \left(\varepsilon_{\mathrm{res}}^{-1} k^3 \log \frac{k}{\varepsilon_{\mathrm{res}}} + \log \frac{3}{\delta} \right) \right\rceil. \end{aligned} \tag{56}$$

The first bound is the sensitivity-sampling bound of Huang and Vishnoi [26, Theorem 3.1(2)]. For the second, the proof of their Lemma 5.11 uses geometric accuracy of order $\varepsilon_{\mathrm{res}}/(z(k+1))$, requiring $O(zk^2 \varepsilon_{\mathrm{res}}^{-1} \log(zk/\varepsilon_{\mathrm{res}}))$ spanning points. The dimension and sampling bounds in the proof of their Theorem 5.10, together with $z \log(zk/\varepsilon_{\mathrm{res}}) = O((1+z)^2 \log(k/\varepsilon_{\mathrm{res}}))$, give the stated dependence on z. Their proofs of Theorems 3.1(2) and 5.10 give these guarantees for every fixed integer sample count at or above the respective threshold. For every fixed integer $s \ge \max\{N_{\mathrm{low}}, N_{\mathrm{res}}, N_{\mathrm{proj}}\}$, a union bound shows that (42), (48), and (49) hold simultaneously with probability at least $1 - \delta$.

When (42), (48), and (49) hold simultaneously, the argument in Part V establishes (43). Part I then gives the simultaneous coreset guarantee when $D > 0$.

Combining the cases $D = 0$ and $D > 0$, define

$$s_0(\xi, \delta; \alpha, k, h, z) := \max\left\{ 1, N_{\mathrm{low}}, N_{\mathrm{res}}, N_{\mathrm{proj}}, \left\lceil c_0 h \xi^{-2} \log \frac{3h}{\delta} \right\rceil \right\}.$$

This choice depends only on $(\xi, \delta; \alpha, k, h, z)$ and dominates the sample-count requirements in both cases. The preceding arguments prove the coreset guarantee for every fixed integer $s \ge s_0$ in either case. To bound this common threshold, substitute $\tau^{-1} = 6\alpha(84z)^{2z} 100^{z+3} \xi^{-(z+3)}$ from (44) into (56) and the definition of N_{proj}. Combining these bounds with the zero-residual term $\lceil c_0 h \xi^{-2} \log(3h/\delta) \rceil$ in the definition of s_0 gives, for a constant $C_z \ge 1$ with $C_z = (1+z)^{O(z)}$,

$$s_0 \le C_z \alpha^{10} k^3 (h+1)^2 \xi^{-(5z+15)} \left(1 + \log \frac{6\alpha^2 k (h+1)}{\xi \delta} \right) + 2.$$

The exponent $5z + 15$ comes from the first term in N_{res} in (56). In this term, the definitions of G_{res} and $\varepsilon_{\mathrm{res}}$ give

$$G_{\mathrm{res}}^2 \varepsilon_{\mathrm{res}}^{-3} = \left(\frac{2\alpha(h+1)}{\tau} \right)^2 \left(\frac{2\alpha}{\tau} \right)^3 = 32\alpha^5 (h+1)^2 \tau^{-5}.$$

The substitution for τ^{-1} therefore gives the factor $\xi^{-5(z+3)} = \xi^{-(5z+15)}$. Ignoring logarithmic factors, all other sample-count terms have smaller powers of ξ^{-1}. The ceilings add at most one to the largest real threshold. The factors depending only on z, including those inside logarithms, are absorbed into C_z. For $\xi = \varepsilon/10$, $\delta = \delta_T = 1/(c(T+1))$, and $h \le H$, the logarithmic factor is at most $\log(120\mathrm{e}c\alpha^2 k(H+1)T/\varepsilon)$ since $T \ge 1$. Thus (34) holds with $L_{\alpha,k,H} = 120\mathrm{e}c\alpha^2 k(H+1)$ and $K_{z,\alpha,k,H} = C_z 10^{5z+15}\alpha^{10}k^3(H+1)^2$. □

M.2 Online low-rank matrix approximation

This section studies online low-rank matrix approximation in the uncentered online principal component analysis model [43, 60].

M.2.1 Problem setup and computational model

Fix an integer $1 \le k \le d$. At each round t, the learner plays a matrix $Z_t \in \mathbb{R}^{d\times k}$ with orthonormal columns, which form a basis of a k-dimensional subspace, and incurs the loss

$$\ell(Z_t, a_t) = \|a_t - Z_t Z_t^\top a_t\|_2^2,$$

where $a_t \in \mathbb{R}^d$ is the t-th column vector. Then, the learner receives a_t as feedback. We assume $\|a_t\|_2^2 \le 1$ for every t, which ensures $\ell(Z, a_t) \in [0,1]$ for every feasible rank-k projection subspace Z. For a prefix $Y = \{a_1, \dots, a_m\}$, write $A_Y = [a_1, \dots, a_m] \in \mathbb{R}^{d\times m}$, and define

$$\mathrm{OPT}(Y) = \min_{Z\in\mathbb{R}^{d\times k}:\, Z^\top Z = I_k} \sum_{a\in Y} \|a - ZZ^\top a\|_2^2.$$

The offline algorithm uses ridge-leverage sampling, following the procedure used by Dong and Yoshida [19].

Computational convention. For the running-time statement below, we augment the exact-real model with Borel-measurable exact spectral oracles. For each matrix used by the procedure, these oracles return a best rank-k approximation and its residual, the required Moore–Penrose inverse, and a k-dimensional left singular subspace associated with the k largest singular values. Each choice among nonunique solutions follows a fixed measurable tie-breaking rule, shared by the algorithm and its output-law analysis. Each oracle call has unit cost. We also allow unit-cost exact evaluation of logarithms on positive arguments and square roots on nonnegative arguments. Thus the running-time statement counts exact-real scalar operations (including these evaluations), random draws, and spectral-oracle calls, and does not assert finite-bit complexity.

M.2.2 Ridge-leverage sampling

Given a prefix matrix $A = [a_1, \dots, a_m] \in \mathbb{R}^{d\times m}$ and an accuracy parameter $\varepsilon \in (0,1]$, the offline algorithm proceeds as follows.

1. Compute the ridge leverage scores of the columns of A with respect to the target rank k. Specifically, writing A_k for a best rank-k approximation to A and $\lambda_k(A) = \|A - A_k\|_F^2/k$, set
$$\tau_i(A) = a_i^\top \left(AA^\top + \lambda_k(A) I_d\right)^\dagger a_i, \qquad i \in [m].$$
Define $M(A) \coloneqq \sum_{j=1}^m \tau_j(A)$. If $M(A) = 0$, terminate and return the fixed deterministic fallback subspace. Otherwise define $p_i(A) \coloneqq \tau_i(A)/M(A)$. This distribution depends on A and k, but not on ε.
2. Fix an absolute constant $c_{\mathrm{lr}} \ge 1$ and draw $N_\varepsilon \sim R_{\mu(\varepsilon)}$, where
$$\mu(\varepsilon) \coloneqq c_{\mathrm{lr}} k\varepsilon^{-2} \log\left(\mathrm{e} + \frac{kT}{\varepsilon}\right) + 2.$$
Because $\mu(\varepsilon) > 2$, every realized count is positive. Conditional on $N_\varepsilon = s$, sample $I_1, \dots, I_s$ independently from $[m]$ according to $p(A)$.
3. For each sampled index $I_h = i$, draw an independent perturbation
$$\widetilde{p}_h \sim \mathrm{Unif}([p_i(A), (1+\varepsilon/2)p_i(A)])$$
and insert the rescaled column $a_i/\sqrt{s\widetilde{p}_h}$ into matrix B.
4. Return a minimizer of $Z \mapsto \|B - ZZ^\top B\|_F^2$, equivalently the top-$k$ left singular subspace of B, using SVD and a fixed deterministic tie-breaking rule.

M.2.3 Approximation and average sensitivity

We establish approximation and average-sensitivity guarantees for the procedure with perturbed weights and a randomly rounded sample count. To prove these guarantees, we use two properties of ridge-leverage sampling: simultaneous preservation of projection costs and a bound on the change in the selected-column law under deletion. The proof of the deletion bound is given in Appendix M.2.6.

Projection-cost preservation for approximation. We use the following integer-count consequence of the proof of Cohen et al. [15, Theorem 4.2], taking each exact ridge leverage score as its own overestimate. For projection-cost preservation, there is an absolute constant $c_{\mathrm{pcp}} > 0$ such that, for every A with $M(A) > 0$, every $\xi, \delta \in (0, 1/2]$, and every integer s satisfying the sample-count threshold

$$s \geq s_{\mathrm{pcp}}(A, \xi, \delta) := \left\lceil c_{\mathrm{pcp}} M(A) \xi^{-2} \log\left(\frac{k}{\delta}\right) \right\rceil, \tag{57}$$

the following holds. Draw $I_1, \ldots, I_s$ independently from $p(A)$ and let $B_s^{(0)} \in \mathbb{R}^{d \times s}$ have h-th column $a_{I_h} / \sqrt{s p_{I_h}(A)}$. With probability at least $1 - \delta$, we have

$$(1-\xi)\left\|(I_d - ZZ^\top)A\right\|_F^2 \leq \left\|(I_d - ZZ^\top)B_s^{(0)}\right\|_F^2 \leq (1+\xi)\left\|(I_d - ZZ^\top)A\right\|_F^2 \tag{58}$$

simultaneously for every $Z \in \mathbb{R}^{d \times k}$ with $Z^\top Z = I_k$. To obtain this integer-count form from the proof of Cohen et al. [15, Theorem 4.2], choose the sample-size multiplier c used in that theorem as $c := s\xi^2 / (M(A) \log(k/\delta))$. Condition (57) gives $c \geq c_{\mathrm{pcp}}$. The matrix concentration and trace-tail estimates in the proofs of Cohen et al. [15, Theorem 4.1 and Lemma A.1] hold for every such c, and the remaining projection-cost comparison in the proof of Cohen et al. [15, Theorem 4.2] is deterministic.

To make this guarantee apply at both possible rounded counts, choose $c_{\mathrm{lr}} \geq 576 c_{\mathrm{pcp}}$ and put $\delta_T := (2T)^{-1}$. When $M(A) > 0$, Lemma 3.1 of Cohen et al. [15] gives $M(A) \leq 2k$. Moreover, we have $\log(k/\delta_T) = \log(2kT) \leq 2\log(\mathrm{e} + kT/\varepsilon)$ because $2kT \leq (\mathrm{e} + kT/\varepsilon)^2$ for $k, T \geq 1$ and $\varepsilon \in (0, 1]$. Hence we have

$$c_{\mathrm{pcp}} M(A) (\varepsilon/12)^{-2} \log\left(\frac{k}{\delta_T}\right) \leq 576 c_{\mathrm{pcp}} k \varepsilon^{-2} \log\left(\mathrm{e} + \frac{kT}{\varepsilon}\right) \leq \mu(\varepsilon) - 2. \tag{59}$$

The additive constant 2 in $\mu(\varepsilon)$ and (59) imply $s_{\mathrm{pcp}}(A, \varepsilon/12, \delta_T) \leq \lfloor \mu(\varepsilon) \rfloor$. By Lemma I.3, every realization of N_ε therefore satisfies (57) with $(\xi, \delta) = (\varepsilon/12, \delta_T)$. When $M(A) = 0$, Step 1 returns the deterministic fallback subspace, and neither $p(A)$ nor (58) is invoked.

Selection stability for average sensitivity. For deletion sensitivity, we first bound the change in the selected-occurrence law at a fixed sample count. Fix a fallback symbol $\perp$ distinct from every occurrence label.

Lemma M.11 (Fixed-size ridge-leverage sensitivity)**.** *Fix a nonempty prefix Y with $|Y| = t$ and a deterministic sample size $s \in \mathbb{Z}_{\geq 0}$. Let $P_{Y,s}^{\mathrm{sel}}$ be the law of the sequence of s selected occurrences drawn from the normalized ridge-leverage distribution of A_Y. For the input $Y \setminus a$, view the law $P_{Y\setminus a,s}^{\mathrm{sel}}$ on the same occurrence labels as Y by assigning zero probability to the removed occurrence. If the ridge leverage scores of the corresponding matrix sum to zero, define the selection law as the point mass on the length-s sequence $(\perp, \ldots, \perp)$. Then, we have*

$$\frac{1}{t} \sum_{a \in Y} \mathrm{TV}\left(P_{Y,s}^{\mathrm{sel}}, P_{Y\setminus a,s}^{\mathrm{sel}}\right) \leq \frac{(4k+5)s}{t}.$$

We defer the lengthy proof to Appendix M.2.6 to avoid interrupting the flow of the argument.

Guarantees for the output subspace. For a prefix Y, if Step 1 does not terminate, let $\mathcal{T}_\varepsilon(Y)$ be the joint law of N_ε and $(I_h, \widetilde{p}_h)_{h=1}^{N_\varepsilon}$ generated from A_Y; if Step 1 terminates, let $\mathcal{T}_\varepsilon(Y)$ be a point mass at a fixed symbol. The deterministic post-processing maps this symbol to the fixed fallback

subspace. Let $\mathcal{L}_\varepsilon(Y)$ denote the output law after deterministic post-processing. When the input matrix is denoted by A, we use the equivalent notation $\mathcal{T}_\varepsilon(A)$ and $\mathcal{L}_\varepsilon(A)$. We also set

$$\varphi(\varepsilon) \coloneqq C_{\mathrm{lr}} k^2 \varepsilon^{-3} \log\left(\mathrm{e} + \frac{kT}{\varepsilon}\right)$$

where C_{lr} is any absolute constant satisfying $C_{\mathrm{lr}} \geq 90(c_{\mathrm{lr}} + 2)$.

The following proposition bounds the expected projection error and the average sensitivity of the subspace returned by the perturbed ridge-leverage procedure.

Proposition M.12 (Fixed-accuracy guarantee for low-rank approximation)**.** *For every fixed $\varepsilon \in (0, 1]$ and every nonempty prefix Y of size t, the randomly rounded perturbed ridge-leverage procedure satisfies*

$$\mathbb{E}_{Z \sim \mathcal{L}_\varepsilon(Y)}\left[\sum_{a \in Y} \left\|a - ZZ^\top a\right\|_2^2\right] \leq (1+\varepsilon)\mathrm{OPT}(Y) + \frac{t}{2T},$$

and its average sensitivity is at most $\varphi(\varepsilon)/(2t)$.

Proof. **Approximation.** If $M(A_Y) = 0$, positivity of the ridge-score quadratic form on every nonzero column implies $A_Y = 0$, and the deterministic fallback subspace has zero loss. Suppose that $M(A_Y) > 0$. Conditional on any realized count $N_\varepsilon = s$, (59) permits the use of (58) with $(\xi, \delta) = (\varepsilon/12, \delta_T)$. The perturbation in Step 3 gives, for every $Z \in \mathbb{R}^{d \times k}$ with $Z^\top Z = I_k$,

$$\frac{\left\|(I_d - ZZ^\top)B_s^{(0)}\right\|_F^2}{1 + \varepsilon/2} \leq \left\|(I_d - ZZ^\top)B\right\|_F^2 \leq \left\|(I_d - ZZ^\top)B_s^{(0)}\right\|_F^2. \tag{60}$$

By (58) and (60), Remark K.2 with $\xi = \varepsilon/12$ and $\beta = \varepsilon/2$ gives an output ratio of at most $1 + \varepsilon$. On the complementary event, every feasible subspace has prefix loss at most t because $\|a\|_2^2 \leq 1$ for every $a \in Y$. Since $\delta_T = 1/(2T)$, averaging over the realized count and the sample gives the stated additive term $t/(2T)$.

Average sensitivity. Retain the original occurrence labels and the fallback symbol $\perp$. Let p_Y and $p_{Y \setminus a}$ be the one-draw selection laws, taking the point mass at $\perp$ for a zero input matrix. Applying Lemma M.11 with $s = 1$ first gives

$$\frac{1}{t}\sum_{a \in Y} \mathrm{TV}(p_Y, p_{Y \setminus a}) \leq \frac{4k+5}{t}. \tag{61}$$

For a fixed count s, let $P_{Y,s}^{\mathrm{den}}$ be the joint law of the selected labels and perturbed probability denominators $(I_h, \widetilde{p}_h)_{h=1}^s$ from Steps 2–3. For a zero matrix, take the point mass on s copies of $(\perp, 0)$; the common output map ignores these pairs and returns the fixed fallback subspace. For two nonzero inputs, apply the first bound in Lemma I.6 with $\mu = 1$ and $\beta = \varepsilon/2$. Because $R_1 = 1$ almost surely and $\varepsilon \leq 1$, the one-draw distance is at most $(5/\varepsilon)\mathrm{TV}(p_Y, p_{Y \setminus a})$. The product-TV inequality therefore gives

$$\mathrm{TV}\left(P_{Y,s}^{\mathrm{den}}, P_{Y \setminus a, s}^{\mathrm{den}}\right) \leq \frac{5s}{\varepsilon}\mathrm{TV}(p_Y, p_{Y \setminus a}).$$

This inequality also covers the fallback cases: with exactly one zero matrix the one-draw selection distance is one, and the distance between the count-s pair laws is one for $s \geq 1$ and zero for $s = 0$; with two zero matrices both distances vanish. Averaging over deletions and using (61) now yields

$$\frac{1}{t}\sum_{a \in Y} \mathrm{TV}\left(P_{Y,s}^{\mathrm{den}}, P_{Y \setminus a, s}^{\mathrm{den}}\right) \leq \frac{5(4k+5)s}{\varepsilon t}.$$

Since deterministic SVD post-processing cannot increase total variation, averaging over $s = N_\varepsilon \sim R_{\mu(\varepsilon)}$ and using $\mathbb{E}[N_\varepsilon] = \mu(\varepsilon)$ gives

$$\frac{1}{t}\sum_{a \in Y} \mathrm{TV}(\mathcal{L}_\varepsilon(Y), \mathcal{L}_\varepsilon(Y \setminus a)) \leq \frac{5(4k+5)\mu(\varepsilon)}{\varepsilon t} \leq \frac{\varphi(\varepsilon)}{2t},$$

where the second inequality follows from $4k + 5 \leq 9k$, $\mu(\varepsilon) \leq (c_{\mathrm{lr}} + 2)k\varepsilon^{-2}\log(\mathrm{e} + kT/\varepsilon)$, and $C_{\mathrm{lr}} \geq 90(c_{\mathrm{lr}} + 2)$. □

M.2.4 Stability under changes in target accuracy

The ridge-leverage distribution is fixed when only the accuracy parameter changes. We compare the rounded counts and perturbation widths, then transfer their total-variation bound through the common SVD output map.

Proposition M.13 (Parameter-change stability for low-rank approximation)**.** *For $0 < \varepsilon \le \eta \le 1$, it holds that*

$$\mathrm{TV}(\mathcal{L}_\varepsilon(A), \mathcal{L}_\eta(A)) \le \mathrm{TV}(\mathcal{T}_\varepsilon(A), \mathcal{T}_\eta(A)) = O\Big(\mu(\varepsilon)\frac{\eta-\varepsilon}{\varepsilon}\Big).$$

Consequently, the randomly rounded ridge-leverage procedure satisfies Assumption 9 with $C_{\mathrm{pc}} = O(1)$.

Proof. If $\sum_i \tau_i(A) = 0$, both transcript laws are the same point mass at the termination symbol and both output laws are the same point mass at the fixed fallback subspace, and hence the claim is immediate. Otherwise, the ridge-leverage distribution $p(A)$ is independent of the accuracy parameter; only the mean $\mu(\varepsilon)$ of N_ε and the perturbation intervals depend on the parameter. Apply Lemma I.7 with $\mathcal{T} = \mathcal{T}^{\mathrm{den}}$, $c = 1/2$, $a = 2$, $C = c_{\mathrm{lr}}k$, and $B = kT$. Using the common measurable SVD output map gives

$$\mathrm{TV}(\mathcal{L}_\varepsilon(A), \mathcal{L}_\eta(A)) \le \mathrm{TV}(\mathcal{T}_\varepsilon(A), \mathcal{T}_\eta(A)) \le 4\mu(\varepsilon)\frac{\eta-\varepsilon}{\varepsilon}.$$

For Assumption 9, fix a dataset X and $x \in X$, let $Y = X \setminus x$, and set $\varepsilon = \varepsilon_X$ and $\eta = \varepsilon_Y$. Since the losses are bounded and non-negative, we have $0 \le \mathrm{OPT}(X) - \mathrm{OPT}(Y) \le 1$ and $\varepsilon_X \le \varepsilon_Y$. By using the above total variation bound for fixed Y and the chosen constant C_{lr}, which makes $\mu(\varepsilon)/\varepsilon = O(\varphi(\varepsilon))$, we obtain

$$\mathrm{TV}(\mathcal{L}_{\varepsilon_X}(Y), \mathcal{L}_{\varepsilon_Y}(Y)) = O(\varphi(\varepsilon_X)(\varepsilon_Y - \varepsilon_X)).$$

By Lemma J.1, the right-hand side is $O(\varepsilon_X^2(\mathrm{OPT}(X) - \mathrm{OPT}(Y)))$; since $\varepsilon_X \le 1$, this is at most $O(\varepsilon_X(\mathrm{OPT}(X) - \mathrm{OPT}(Y)))$. Averaging over $x \in X$ yields the condition with $C_{\mathrm{pc}} = O(1)$. □

M.2.5 Regret and oracle complexity

By Propositions M.12 and M.13, the low-rank approximation offline algorithm family satisfies Assumption 4 with $\varphi(\varepsilon) = C_{\mathrm{lr}}k^2\varepsilon^{-3}\log(\mathrm{e} + kT/\varepsilon)$ and Assumption 9 with $C_{\mathrm{pc}} = O(1)$. Thus, applying Corollary 12 with $q = 3$, $\lambda = H_T^{-(q+1)/q}$, $C_1 = C_{\mathrm{lr}}k^2$, $C_2 = kT$, and $C_{\mathrm{pc}} = O(1)$ gives

$$\mathcal{R}_T = \tilde{O}\Big(k^2 + k^{1/2}\Big(1 + \mathrm{OPT}_T^{3/4}\Big)\Big).$$

To bound the per-round cost uniformly over the online history and sampling randomness, put $\varepsilon_\star := \varepsilon_{\min}(T + \lambda)$. Lemma I.4 gives the deterministic bound

$$N_{\varepsilon_t} \le \lceil \mu(\varepsilon_\star) \rceil \quad \text{almost surely.}$$

Comparing the objective defining $\varepsilon_\star$ with the feasible point $\varepsilon = 1$ gives $\varepsilon_\star^{-3} \le (T + \lambda)/(C_{\mathrm{lr}}k^2) + \log(\mathrm{e} + kT)$. Hence the bound on N_{ε_t} is polynomial in k and T. A constant number of calls to the unit-cost exact spectral oracles gives A_k, the Moore–Penrose inverse in $\tau_i(A)$, the exact prefix optimum, and the returned top-k subspace selected by the fixed tie-breaking rule. Since $m \le T$, forming all scores, generating the rounded sample, and assembling B require polynomially many exact-real scalar operations and random draws. Thus the adaptive low-rank procedure has polynomial worst-case per-prefix oracle-operation cost.

M.2.6 Proof of the ridge-leverage deletion bound

We prove Lemma M.11 by first comparing one sampled column before and after deletion. Deletion changes both the matrix and its ridge parameter. We bound these two effects separately, sum over the deleted columns, and then use the product total-variation inequality for a sample of fixed size.

Proof of Lemma M.11. Write $A \coloneqq [a_1, \dots, a_t]$, let A_{-i} omit occurrence a_i, and put

$$E \coloneqq \|A - A_k\|_F^2, \qquad E_i \coloneqq \|A_{-i} - (A_{-i})_k\|_F^2, \qquad \Delta_i \coloneqq E - E_i.$$

For every nonzero matrix B, let p_B denote the distribution obtained by normalizing the ridge leverage scores defined in Step 1 for B. We place all laws on the occurrence labels of A together with a fallback label $\perp$: an absent occurrence has probability zero, and the zero matrix uses the point mass at $\perp$. The product-TV inequality in Lemma H.2 gives

$$\frac{1}{t}\sum_{i=1}^{t} \mathrm{TV}\left(p_A^{\otimes s}, p_{A_{-i}}^{\otimes s}\right) \le \frac{s}{t}\sum_{i=1}^{t} \mathrm{TV}(p_A, p_{A_{-i}}). \tag{62}$$

Suppose first that $E > 0$ and set $\gamma \coloneqq E/k$. For a fixed ridge $\gamma > 0$, write p_B^γ for the distribution obtained by normalizing the scores $b_j^\top (BB^\top + \gamma I)^{-1} b_j$. Then we have $p_A = p_A^\gamma$. Since $E > 0$ implies $\mathrm{rank}(A) > k$, deleting one column cannot produce the zero matrix, and hence we have

$$\mathrm{TV}(p_A, p_{A_{-i}}) \le \mathrm{TV}(p_A^\gamma, p_{A_{-i}}^\gamma) + \mathrm{TV}(p_{A_{-i}}^\gamma, p_{A_{-i}}). \tag{63}$$

To analyze the first term, put $G \coloneqq AA^\top + \gamma I$, $\tau_j \coloneqq a_j^\top G^{-1} a_j$, and $Q \coloneqq \sum_j \tau_j$. Let $\widehat{\tau}_j, \widehat{Q}$ be the corresponding quantities after deleting a_i while retaining the same γ. Since $G - a_i a_i^\top \succ 0$, we have $\tau_i < 1$. Sherman–Morrison and $AA^\top = G - \gamma I$ give, for $j \neq i$,

$$\widehat{\tau}_j = \tau_j + \frac{(a_j^\top G^{-1} a_i)^2}{1 - \tau_i} \ge \tau_j, \qquad \widehat{Q} = Q - \frac{\gamma \|G^{-1} a_i\|_2^2}{1 - \tau_i} \le Q.$$

Thus every surviving occurrence receives at least its former normalized probability. The total increase on surviving occurrences equals the deleted mass. Although $\gamma = E/k$ is fixed here, this calculation uses only $\gamma > 0$ and, whenever both matrices are nonzero, gives

$$\mathrm{TV}(p_A^\gamma, p_{A_{-i}}^\gamma) = (p_A^\gamma)_i. \tag{64}$$

For the current choice $\gamma = E/k$, we have $p_A^\gamma = p_A$.

Next compare the two ridge parameters on A_{-i}. Restricting a best rank-k approximation of A to A_{-i} gives $0 \le E_i \le E$. Let $\delta_i \coloneqq \Delta_i / E$ and $M \coloneqq A_{-i} A_{-i}^\top$. If $\delta_i < 1/2$, then we have $E_i > 0$ and

$$(M + \gamma I)^{-1} \preceq (M + (E_i/k) I)^{-1} \preceq (1 - \delta_i)^{-1} (M + \gamma I)^{-1}.$$

Normalizing the two score vectors therefore gives $\mathrm{TV}(p_{A_{-i}}^\gamma, p_{A_{-i}}) \le (1 - \delta_i)^{-1} - 1 \le 2\delta_i$. If $\delta_i \ge 1/2$, the same conclusion follows from $\mathrm{TV}(p_{A_{-i}}^\gamma, p_{A_{-i}}) \le 1 \le 2\delta_i$. Combining this with (63) and (64) yields

$$\mathrm{TV}(p_A, p_{A_{-i}}) \le (p_A)_i + 2\frac{\Delta_i}{E}. \tag{65}$$

To sum (65) over deletions, we bound the relative decrease Δ_i / E in terms of the deleted column's ridge leverage score. Let P_i be a rank-k orthogonal projector minimizing $\|(I - P) A_{-i}\|_F^2$, and put $r_i \coloneqq \|(I - P_i) a_i\|_2^2$ and $D_i \coloneqq \|(I - P_i) A\|_F^2 = E_i + r_i$. Optimality of A_k gives $E \le D_i$, equivalently $\Delta_i \le r_i$. Since $E = E_i + \Delta_i$, $D_i = E_i + r_i$, and $E_i \ge 0$, we have

$$\frac{\Delta_i}{E} = 1 - \frac{E_i}{E} \le 1 - \frac{E_i}{D_i} = \frac{r_i}{D_i}. \tag{66}$$

Inserting $G^{1/2} G^{-1/2}$ and applying Cauchy–Schwarz gives

$$r_i \le \|(I - P_i) G^{1/2}\|_2^2 \, \tau_i.$$

Moreover, we have

$$\|(I - P_i) G^{1/2}\|_2^2 \le \|(I - P_i) A\|_2^2 + \gamma \le D_i + \frac{D_i}{k} = (1 + 1/k) D_i,$$

where we used $\|(I - P_i) A\|_2^2 \le D_i$ and $k\gamma = E \le D_i$. Together with (66), this proves

$$\frac{\Delta_i}{E} \le (1 + 1/k) \tau_i. \tag{67}$$

The ridge leverage scores τ_i satisfy $\sum_i \tau_i \le 2k$ [15, Lemma 3.1]. Summing (65) and using (67) therefore gives

$$\sum_{i=1}^{t} \mathrm{TV}(p_A, p_{A_{-i}}) \le 1 + 2(1 + 1/k)\, 2k = 4k + 5. \tag{68}$$

It remains to consider $E = 0$. Every nonzero submatrix then has rank at most k. For every such submatrix B, the scores in Step 1 are ordinary leverage scores, and hence p_B is their normalized distribution. For fixed nonzero B, the law p_B^γ converges in total variation to p_B as $\gamma \downarrow 0$: every score converges to $b_j^\top (BB^\top)^\dagger b_j$, and their sum converges to $\mathrm{rank}(B) > 0$. Taking this limit in (64) gives $\mathrm{TV}(p_A, p_{A_{-i}}) = (p_A)_i$ whenever both matrices are nonzero. If deleting a_i produces the zero matrix, both sides equal one; if $A = 0$, all laws are the fallback point mass. Hence the sum in (68) is at most one in this case.

Substituting (68), which holds in both cases, into (62) completes the proof. □

N $\Omega(\sqrt{n\mathrm{OPT}_T})$ lower bound for online submodular function minimization

This section proves the $\Omega(\sqrt{n\mathrm{OPT}_T})$ lower bound for online submodular function minimization mentioned in Section 8, even for a fixed multiset of losses presented in uniformly random order. Below, for a fixed multiset M of T submodular loss functions, write

$$\mathcal{R}_T(M) \coloneqq \mathbb{E}_{\pi,\mathsf{Alg}}\left[\sum_{t=1}^{T} \ell_{\pi(t)}(S_t)\right] - \mathrm{OPT}(M), \qquad \text{where} \qquad \mathrm{OPT}(M) \coloneqq \min_{S \subseteq V} \sum_{\ell_t \in M} \ell_t(S).$$

Theorem N.1. *There exists a universal constant $c > 0$ such that, for every randomized online learner for online submodular function minimization on a ground set of size $1 \le n \le T$, there is a fixed multiset M of T submodular losses such that, when M is revealed in a uniformly random order, it holds that*

$$\mathrm{OPT}(M) \le T/2 \qquad \textit{and} \qquad \mathcal{R}_T(M) \ge c\sqrt{nT}.$$

In particular, $\mathcal{R}_T(M) \ge c\sqrt{n\mathrm{OPT}(M)}$.

Proof. We use the following instance construction of Hazan and Kale [25, Theorem 14] for establishing the $\Omega(\sqrt{nT})$ lower bound for online submodular function minimization.

Let $V \coloneqq [n]$, where $n \le T$. Choose integers $m_i \in \{\lfloor T/n \rfloor, \lceil T/n \rceil\}$ such that $\sum_{i=1}^n m_i = T$. For each coordinate $i \in [n]$ and each copy $j \in [m_i]$, draw an independent Rademacher sign $\sigma_{i,j} \in \{-1, +1\}$ and define the submodular loss

$$\ell_{i,j}(S) = \begin{cases} \frac{1}{2}(1 + \sigma_{i,j}) & \text{if } i \in S, \\ \frac{1}{2}(1 - \sigma_{i,j}) & \text{if } i \notin S. \end{cases}$$

Let M_σ be the resulting multiset. Fix an enumeration $(\ell_1, \dots, \ell_T)$ of M_σ and reveal $\ell_{\pi(t)}$ at time t, where π is a uniformly random permutation of $[T]$. For any online learner, the sign of the next revealed loss is independent of the learner's current choice, and therefore we have

$$\mathbb{E}_{\sigma,\pi,\mathsf{Alg}}\left[\sum_{t=1}^{T} \ell_{\pi(t)}(S_t)\right] = \frac{T}{2},$$

where π denotes the random order and Alg denotes the learner's internal randomness. On the other hand, for each fixed sign realization σ, by including i in S if and only if $\sum_{j=1}^{m_i} \sigma_{i,j} \le 0$, we can achieve

$$\mathrm{OPT}(M_\sigma) = \frac{1}{2}\left(T - \sum_{i=1}^{n} \left|\sum_{j=1}^{m_i} \sigma_{i,j}\right|\right),$$

and thus $\mathrm{OPT}(M_\sigma) \le T/2$ always holds. By Khinchine's inequality, there is an absolute constant $c_1 > 0$ such that

$$\begin{aligned}\mathbb{E}_\sigma[\mathcal{R}_T(M_\sigma)] &= \mathbb{E}_{\sigma,\pi,\mathsf{Alg}}\left[\sum_{t=1}^{T} \ell_{\pi(t)}(S_t)\right] - \mathbb{E}_\sigma[\mathrm{OPT}(M_\sigma)] \\ &= \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_\sigma\left[\left|\sum_{j=1}^{m_i} \sigma_{i,j}\right|\right] \\ &\ge c_1 \sum_{i=1}^{n} \sqrt{m_i} \ge c_1 n \sqrt{\lfloor T/n \rfloor} = \Omega\Big(\sqrt{nT}\Big).\end{aligned}$$

The final relation uses $\lfloor T/n \rfloor \ge T/(2n)$, since $n \le T$. Choose a universal constant $c > 0$ such that the preceding estimate gives $\mathbb{E}_\sigma[\mathcal{R}_T(M_\sigma)] \ge c\sqrt{nT}$. Since the sign space is finite, some realization σ satisfies $\mathcal{R}_T(M_\sigma) \ge c\sqrt{nT}$. Together with $\mathrm{OPT}(M_\sigma) \le T/2$, this proves the theorem. □

Declaration of LLM usage

We used OpenAI's ChatGPT and Codex for manuscript editing, literature checks, and proof development and checking. The authors take responsibility for all content.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See the Abstract and Sections 1 and 1.1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 8.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Sections 2–7 and Appendices A–N.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not contain empirical experiments; its main claims are theoretical and are supported by formal statements and proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not report experimental results whose reproducibility depends on releasing data or code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include training/test experiments or other empirical evaluations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include empirical experiments, so error bars and statistical-significance analyses are not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include empirical experiments whose reproduction would require reporting compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is a theoretical study of online learning under the random-order model and does not involve human subjects, personal data, or other sensitive experimental procedures.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is a foundational theoretical study of online learning algorithms and does not make application-specific deployment claims that would require a dedicated broader-impact discussion in the submission.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The submission does not release a high-risk model, dataset, or similar asset that would require misuse safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use external datasets, pretrained models, or third-party software artifacts as research assets; prior algorithms and theoretical results used as technical ingredients are cited in the bibliography.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, model, or software artifact whose documentation is part of the submission.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our use of ChatGPT and Codex is disclosed in Declaration of LLM usage.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.