Inference for Multiple Quantiles under Local Differential Privacy

Qirui Hu * **Yi Liu** †

Abstract

In this paper, we study the problem of simultaneous inference for a fixed finite collection of quantiles under local differential privacy (LDP). We develop a general online framework that reduces the multiple-quantile problem to a categorical one-hot estimation problem. A projected Robbins-Monro recursion guarantees non-crossing at every iteration, while Polyak-Ruppert averaging together with multivariate self-normalization yields a joint asymptotically pivotal ellipsoidal confidence region without requiring estimation of nuisance densities or asymptotic covariance matrices. We also present several important examples in detail, including direct $(k+1)$-ary randomized response (kRR), optimized unary encoding (OUE), optimized local hashing (OLH), and Hadamard response (HR). Large-sample kRR experiments support the predicted root-n scaling and covariance calibration of the averaged estimator, while showing that self-normalized coverage improves more slowly and remains mildly below the nominal level in finite samples.

1 Introduction

Quantiles are fundamental summaries of a distribution. Unlike the mean, they describe location at arbitrary probability levels, are robust to heavy tails and outliers, and naturally capture distributional spread and tail behavior. In applications, one rarely needs a single quantile in isolation: lower and upper prediction limits, inter-quantile ranges, and distributional summaries all require several quantiles to be estimated jointly. A joint treatment is also structurally desirable, because true quantiles are monotone in the level and separate estimation can lead to crossing estimates; see [20] for background on quantile methodology and [14, 22] for recent private multi-quantile problems in the trusted-curator setting.

Differential privacy (DP) provides a rigorous framework for limiting the information revealed about any single individual through a randomized data-release procedure [11]. In the local differential privacy (LDP) model, each user privatizes their own data before transmission, thereby removing the need for a trusted curator. This model is rooted in randomized response [33] and was formalized in modern learning settings by [19]. It is now a central paradigm for privacy-preserving data collection at scale. The local model is statistically more demanding than the trusted-curator model, which makes careful mechanism design and efficient inference particularly important [9, 6].

Quantiles have been studied under both curator-model DP and LDP. In the curator model, early work used smooth sensitivity and the propose-test-release paradigm to privatize robust statistics such as the median and other quantiles [26, 10]. More recent work developed dedicated algorithms for one or many quantiles, including simultaneous exact-quantile release via the exponential mechanism [14], recursive approximate-quantile algorithms [18], bounded-space methods [3], analyses tailored to distributions with atoms [23], and private estimation of many quantiles [22]. Very recent work

*School of Statistics and Data Science, Shanghai University of Finance and Economics. Institute of Big Data Research, Shanghai University of Finance and Economics. Department of Mathematics, Ruhr University Bochum.

†York University. Corresponding author: yliu42@yorku.ca

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

further sharpened error guarantees for private multi-quantile release [16]. In contrast, quantile-specific inference under curator-model DP remains relatively limited; existing results mainly concern confidence intervals for the median or more generic DP confidence-interval constructions rather than joint multi-quantile inference [8, 12].

Under LDP, direct quantile estimation has followed several routes. Early approaches used interactive binary-search ideas or embedded quantile subroutines inside other tasks [7, 13], while general minimax LDP estimation theory also contains a stochastic-gradient procedure for the median [9]. More recently, [1] developed adaptive distributed protocols with improved sample complexity for single-quantile estimation on discrete domains. Another route is indirect: one may first estimate a private cumulative distribution function and then invert it to recover quantiles [25]. For inference, [24] established online self-normalized inference for a single quantile, and [5] recently studied federated single-quantile inference under LDP. However, these results do not address the problem considered here: joint estimation and joint inference for a fixed collection of quantiles from one private report per user while respecting the natural noncrossing constraint.

The many-quantile problem under LDP is challenging for several reasons. First, estimating each target level separately either requires multiple reports per user or sacrifices efficiency by splitting users or privacy budget across levels. Second, separate estimators ignore the strong dependence among nearby quantiles and can cross in finite samples. Third, classical quantile inference depends on nuisance densities $f\{Q(\tau_j)\}$, while joint LDP inference additionally involves an oracle-specific asymptotic covariance structure induced by the local mechanism. A useful method should therefore update all quantiles simultaneously, preserve monotonicity, and deliver a joint confidence region without repeatedly tuning or estimating difficult nuisance quantities.

Our main contributions are as follows.

- We formulate an online multi-quantile LDP framework based on a generic class of unbiased bounded one-hot decoders. The framework covers several standard categorical mechanisms used in local frequency estimation, including direct $(k+1)$-ary randomized response (kRR), optimized unary encoding (OUE), optimized local hashing (OLH), and Hadamard response (HR) [32, 2, 6]. It allows all target quantiles to be estimated jointly from one privatized report per user.
- We propose a projected stochastic approximation algorithm that updates all k quantiles simultaneously and guarantees noncrossing at every iteration. Under mild conditions, we establish strong consistency, eventual inactivity of the projection, an asymptotic linear representation, a joint central limit theorem for the averaged estimator, and a multivariate self-normalized pivotal limit. The resulting confidence region is an ellipsoid with asymptotically correct joint coverage and does not require estimating $f\{Q(\tau_j)\}$ or the asymptotic covariance matrix.
- We verify the generic oracle conditions for kRR, OUE, OLH, and HR. For kRR and OUE we derive closed-form expressions for the asymptotic prefix covariance matrix. For OLH and HR we provide explicit second-moment formulas and decoder bounds that can be inserted directly into the generic theory.

The paper is organized as follows. Section 2 introduces the model, the generic oracle framework, and the online algorithm. Section 3 states the main theoretical results and presents the concrete mechanisms. Section 4 reports simulations. Section 5 concludes. Appendix B contains formal proofs of all stated results.

2 Methodology

We first recall the definitions of DP and LDP.

Definition 1 (Differential privacy [11])**.** *A randomized algorithm $\mathcal{A}$, taking a dataset consisting of individuals as its input, is (ϵ, δ)-differentially private if, for any pair of datasets S and S' that differ in the record of a single individual and any event E, it satisfies the following condition:*

$$\mathbb{P}[\mathcal{A}(S) \in E] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}\left(S'\right) \in E\right] + \delta.$$

When $\delta = 0$, $\mathcal{A}$ is called ϵ-differentially private (ϵ-DP).

Definition 2 (Local differential privacy [17])**.** *A local randomizer $R : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-locally differentially private if, for all $x, x' \in \mathcal{X}$ and all measurable $A \subseteq \mathcal{Y}$,*

$$\mathbb{P}\{R(x) \in A\} \leq e^{\epsilon}\mathbb{P}\{R(x') \in A\} + \delta.$$

When $\delta = 0$, R is called ϵ-locally differentially private (ϵ-LDP).

Problem setup. Let $X_1, X_2, \ldots$ be i.i.d. real-valued random variables with distribution function F. For $\tau \in (0,1)$ define

$$Q(\tau) := F^{-1}(\tau) := \inf\{x \in \mathbb{R} : F(x) \geq \tau\}.$$

Fix target levels

$$0 < \tau_1 < \cdots < \tau_k < 1, \qquad \boldsymbol{Q} = (Q_1, \ldots, Q_k)^\top, \qquad Q_j := Q(\tau_j).$$

Throughout the paper, k and $\varepsilon > 0$ are fixed as $n \to \infty$.

Assumption 1. *The variables X_t are i.i.d. on $\mathbb{R}$ with distribution function F. There exist known constants $L < U$ such that $L < Q_1 < \cdots < Q_k < U$, and for each j, Q_j is the unique solution of $F(q) = \tau_j$ in $[L, U]$. Moreover, F is continuously differentiable in neighborhoods of $Q_1, \ldots, Q_k$, with density $f = F'$, and there exist constants $0 < \underline{f} \leq \overline{f} < \infty$ such that $\underline{f} \leq f(x) \leq \overline{f}$ for all x in those neighborhoods.*

Assumption 1 is standard in quantile estimation and inference. The boundedness of the target quantiles avoids additional complications associated with extreme quantiles, while the existence and boundedness of the density function in a neighborhood of the target values ensure that the asymptotic covariance matrix is well defined. The box must contain all target quantiles strictly in its interior. Wider valid bounds can lengthen transients; bounds that exclude a target may lead to a constrained boundary target, invalidating the stated confidence region for the true quantiles. Bounds should be fixed from public information; data-dependent selection requires separate privacy accounting.

Define the monotone parameter set $\mathcal{C} := \{\boldsymbol{q} \in [L, U]^k : q_1 \leq \cdots \leq q_k\}$. Projection onto $\mathcal{C}$ is Euclidean isotonic regression and can be computed by the pool-adjacent-violators algorithm [4].

At time t, the server holds the current monotone estimate $\boldsymbol{q}_{t-1} = (q_{t-1,1}, \ldots, q_{t-1,k})^\top \in \mathcal{C}$. User t converts their observation into the interval label

$$C_t(\boldsymbol{q}_{t-1}) := \sum_{j=1}^{k} \mathbf{1}\{X_t > q_{t-1,j}\} \in \{0, 1, \ldots, k\}. \tag{1}$$

Thus the multiple-quantile problem is reduced to a categorical problem on the alphabet $\{0, 1, \ldots, k\}$.

Let $\boldsymbol{e}_r \in \mathbb{R}^{k+1}$ denote the one-hot vector of category r. For any threshold vector $\boldsymbol{q} \in \mathcal{C}$, define the interval-probability vector $\pi(\boldsymbol{q}) = (\pi_0(\boldsymbol{q}), \ldots, \pi_k(\boldsymbol{q}))^\top$, where

$$\pi_0(\boldsymbol{q}) = F(q_1), \quad \pi_r(\boldsymbol{q}) = F(q_{r+1}) - F(q_r)\ (1 \leq r \leq k-1), \quad \pi_k(\boldsymbol{q}) = 1 - F(q_k). \tag{2}$$

Then $C_t(\boldsymbol{q}) \sim \pi(\boldsymbol{q})$.

Generic one-hot oracle. We now state a generic condition on the local mechanism.

Assumption 2 (Generic one-hot oracle)**.** *There exist a local randomizer $\mathcal{R} : \{0, 1, \ldots, k\} \to \mathcal{Y}$ and a decoder $\phi = (\phi_0, \ldots, \phi_k) : \mathcal{Y} \to \mathbb{R}^{k+1}$ such that:*

(i) The randomizer $\mathcal{R}$ is ε-LDP;

(ii) if $Y \sim \mathcal{R}(c)$, then $\mathbb{E}\big[\phi(Y)\big] = \boldsymbol{e}_c$;

(iii) there exists a finite constant $B_{\varepsilon,k}$ such that $\sup_{c \in \{0,\ldots,k\}} \|\phi\{\mathcal{R}(c)\}\|_\infty \leq B_{\varepsilon,k}$ almost surely.

Assumption 2(ii) says that the decoder is unbiased for the one-hot representation of the interval label. Assumption 2(iii) yields a uniform boundedness condition that greatly simplifies the stochastic approximation and self-normalization arguments.

Define the prefix-sum matrix $P \in \mathbb{R}^{k \times k+1}$ by

$$P_{j,r} = \mathbf{1}\{r \leq j-1\}, \qquad j = 1, \ldots, k,\ r = 0, \ldots, k.$$

If $v = \boldsymbol{e}_c$, then $(Pv)_j = \mathbf{1}\{c \le j-1\}$, i.e., whether the category lies at or below threshold q_j. Therefore, after receiving the private report Y_t, the server forms

$$\widehat{v}_t := \phi(Y_t) \in \mathbb{R}^{k+1}, \qquad \widehat{G}_t := P\widehat{v}_t \in \mathbb{R}^k.$$

The jth coordinate $\widehat{G}_{j,t}$ is an unbiased estimate of $F(q_{t-1,j})$.

For a generic oracle $\mathcal{O} = (\mathcal{R}, \phi)$ satisfying Assumption 2, define the oracle-specific conditional second-moment matrices

$$M_r^{\mathcal{O}} := \mathbb{E}\{\phi(Y)\phi(Y)^\top \mid C = r\} \in \mathbb{R}^{k+1\times k+1}, \qquad r = 0, \dots, k.$$

Then the unconditional decoded one-hot second moment, covariance, and prefix covariance at threshold vector $\boldsymbol{q}$ are

$$M_{\mathcal{O}}(\boldsymbol{q}) := \sum_{r=0}^{k} \pi_r(\boldsymbol{q}) M_r^{\mathcal{O}}, \qquad \Gamma_{\mathcal{O}}(\boldsymbol{q}) := M_{\mathcal{O}}(\boldsymbol{q}) - \pi(\boldsymbol{q})\pi(\boldsymbol{q})^\top, \qquad \Sigma_{\mathcal{O}}(\boldsymbol{q}) := P\Gamma_{\mathcal{O}}(\boldsymbol{q})P^\top.$$

The asymptotic quantile covariance will turn out to be driven by $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$.

LDP estimation and self-normalized inference. To estimate the quantiles and construct the confidence regions, we propose an online algorithm with stochastic approximation and self-normalization. Define

$$G(\boldsymbol{q}) := (F(q_1), \dots, F(q_k))^\top, \qquad h(\boldsymbol{q}) := \boldsymbol{\tau} - G(\boldsymbol{q}), \qquad \boldsymbol{\tau} = (\tau_1, \dots, \tau_k)^\top.$$

Then $h(\boldsymbol{Q}) = 0$. Given a step-size sequence $\gamma_t = \gamma_0 t^{-a}, a \in (1/2, 1),\ \gamma_0 > 0$, our update is the projected Robbins–Monro recursion

$$\boldsymbol{q}_t = \Pi_{\mathcal{C}}\big\{\boldsymbol{q}_{t-1} + \gamma_t(\boldsymbol{\tau} - \widehat{G}_t)\big\}. \tag{3}$$

Projection enforces noncrossing exactly. The averaged estimator is $\bar{\boldsymbol{q}}_t := t^{-1}\sum_{s=1}^{t} \boldsymbol{q}_s$.

For joint inference define the multivariate self-normalizer

$$W_n := \frac{1}{n^2}\sum_{t=1}^{n} t^2(\bar{\boldsymbol{q}}_t - \bar{\boldsymbol{q}}_n)(\bar{\boldsymbol{q}}_t - \bar{\boldsymbol{q}}_n)^\top. \tag{4}$$

For any candidate parameter $\theta \in \mathbb{R}^k$ define

$$T_n(\theta) := n(\bar{\boldsymbol{q}}_n - \theta)^\top W_n^{-1}(\bar{\boldsymbol{q}}_n - \theta). \tag{5}$$

The statistic is evaluated on the event that W_n is nonsingular with probability tending to one. Let $\mathbf{B}(\cdot)$ be a standard k-dimensional Brownian motion and define

$$U_k := \mathbf{B}(1)^\top V^{-1}\mathbf{B}(1), \qquad V := \int_0^1 (\mathbf{B}(r) - r\mathbf{B}(1))(\mathbf{B}(r) - r\mathbf{B}(1))^\top dr. \tag{6}$$

If $u_{k,1-\alpha}$ is the $(1-\alpha)$ quantile of U_k, the joint self-normalized $(1-\alpha)$ confidence region is

$$\mathcal{R}_{n,1-\alpha} := \{\theta \in \mathbb{R}^k : T_n(\theta) \le u_{k,1-\alpha}\}.$$

This is an ellipsoid whenever W_n is invertible.

3 Theoretical results

We first establish privacy, unbiasedness, consistency, and the eventual inactivity of the projection step. We then derive the oracle-dependent asymptotic law of the averaged estimator and a self-normalized joint confidence region. The mechanism enters the limit theory only through the matrix $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$.

Theorem 1 (Privacy, unbiasedness, and noncrossing)**.** *Suppose Assumption 2 holds.*

(i) Conditional on the server query $\boldsymbol{q}_{t-1}$, the mechanism mapping X_t to Y_t is ε-LDP. Hence Algorithm 1 is per-user ε-LDP.

Algorithm 1 Online joint quantile inference under a generic one-hot oracle

Require: privacy level ε, target levels $0 < \tau_1 < \cdots < \tau_k < 1$, box $L < U$, step sizes $\gamma_t = \gamma_0 t^{-a}$

initialize $\boldsymbol{q}_0 \in \mathcal{C}$ and $\bar{\boldsymbol{q}}_0 = \boldsymbol{q}_0$
for $t = 1, 2, \ldots, n$ **do**
    server broadcasts $\boldsymbol{q}_{t-1}$
    user computes C_t from (1)
    user draws $Y_t \sim \mathcal{R}(C_t)$ and sends Y_t
    server decodes $\widehat{v}_t = \phi(Y_t)$ and $\widehat{G}_t = P\widehat{v}_t$
    update $\boldsymbol{q}_t = \Pi_{\mathcal{C}}\{\boldsymbol{q}_{t-1} + \gamma_t(\boldsymbol{\tau} - \widehat{G}_t)\}$
    update average $\bar{\boldsymbol{q}}_t = ((t-1)/t)\bar{\boldsymbol{q}}_{t-1} + \boldsymbol{q}_t/t$
end for
form W_n from (4) and the ellipsoid $\mathcal{R}_{n,1-\alpha}$ using the statistic in (5)
return $\bar{\boldsymbol{q}}_n$ and $\mathcal{R}_{n,1-\alpha}$

(ii) For every $t \geq 1$, $\mathbb{E}[\widehat{G}_t \mid \boldsymbol{q}_{t-1}] = G(\boldsymbol{q}_{t-1})$.

(iii) For every $t \geq 0$, $\boldsymbol{q}_t \in \mathcal{C}$. Since $\mathcal{C}$ is convex, $\bar{\boldsymbol{q}}_t \in \mathcal{C}$ for every $t \geq 1$.

The next theorem gives consistency. Standard projected stochastic-approximation arguments are based on supermartingale convergence and recursive-algorithm theory; see, for example, [29, 21].

Assumption 3 (Step sizes)**.** *The step sizes satisfy $\gamma_t = \gamma_0 t^{-a}$ with $a \in (1/2, 1)$ and $\gamma_0 > 0$.*

Theorem 2 (Strong consistency)**.** *Suppose Assumptions 1–3 hold. Then*

$$\boldsymbol{q}_t \xrightarrow{a.s.} \boldsymbol{Q}, \qquad \bar{\boldsymbol{q}}_t \xrightarrow{a.s.} \boldsymbol{Q}.$$

Projection is crucial in finite samples but asymptotically inactive because the true quantile vector lies in the interior of $\mathcal{C}$.

Theorem 3 (Asymptotic irrelevance of projection)**.** *Under the assumptions of Theorem 2, there exists an almost surely finite random time T such that for all $t \geq T$, $\boldsymbol{q}_t = \boldsymbol{q}_{t-1} + \gamma_t(\boldsymbol{\tau} - \widehat{G}_t)$, that is, the projection in* (3) *is inactive from time T onward.*

We now state the oracle-dependent asymptotic law. Define $D := \operatorname{diag}\{f(Q_1), \ldots, f(Q_k)\}$. Since $F(Q_j) = \tau_j$, Assumption 1 implies that D is positive definite. Let $\Omega_{\mathcal{O}} := D^{-1}\Sigma_{\mathcal{O}}(\boldsymbol{Q})D^{-1}$.

Theorem 4 (Asymptotic linear representation and joint CLT)**.** *Suppose Assumptions 1–3 hold, $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$ is positive definite, and let $\mathcal{O} = (\mathcal{R}, \phi)$ satisfy Assumption 2. Define $M_t = G(\boldsymbol{q}_{t-1}) - \widehat{G}_t$. Then*

$$\sqrt{n}(\bar{\boldsymbol{q}}_n - \boldsymbol{Q}) = D^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} M_t + o_p(1).$$

Moreover, the conditional covariance of M_t converges almost surely to $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$. Consequently,

$$\sqrt{n}(\bar{\boldsymbol{q}}_n - \boldsymbol{Q}) \xrightarrow{d} \mathcal{N}(0, \Omega_{\mathcal{O}}).$$

Equivalently, if $\xi := \boldsymbol{\tau} - P\phi(Y^*)$ with $Y^* \sim \mathcal{R}(C^*)$ and $C^* \sim \pi(\boldsymbol{Q})$, then $\operatorname{Var}(\xi) = \Sigma_{\mathcal{O}}(\boldsymbol{Q})$.

Theorem 4 isolates all oracle dependence in $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$ and therefore in $\Omega_{\mathcal{O}}$. At the proof level, the argument combines stochastic-approximation linearization with a martingale central limit theorem; see [15].

Corollary 1 (Oracle comparison criterion)**.** *Let Ψ be any monotone matrix-risk functional, for example $\Psi(M) = \operatorname{tr}(M)$ or $\Psi(M) = \log\det(M)$, when defined. Let $\mathfrak{O}_\varepsilon$ be a class of ε-LDP one-hot oracles satisfying Assumption 2. If the minimum is attained, any oracle*

$$\mathcal{O}^\star_\Psi \in \operatorname*{argmin}_{\mathcal{O} \in \mathfrak{O}_\varepsilon} \Psi\big(\Omega_{\mathcal{O}}\big)$$

is asymptotically optimal for the averaged quantile estimator within the class $\mathfrak{O}_\varepsilon$ under criterion Ψ.

Corollary 1 reduces fixed-k mechanism design to an oracle-comparison problem. The recursion adds no further oracle-specific term to the limiting covariance. Trace risk uses the density-weighted $\Omega_{\mathcal{O}}$; this criterion is not a data-driven selection rule or a claim of efficiency over all LDP procedures. In regular parametric scalar models, efficient lower and upper bounds are now known under LDP [31]. For semiparametric multiple-quantile inference, however, a complete efficiency characterization remains open.

The next theorem yields a nuisance-free joint confidence region. The self-normalized statistic follows the same broad principle as fixed-dimensional self-normalization in time-series and dependent-data inference [30].

Theorem 5 (Multivariate self-normalization)**.** *Under the assumptions of Theorem 4,*

$$T_n(\boldsymbol{Q}) \xrightarrow{d} U_k,$$

where U_k is defined in (6)*. Consequently, $\mathbb{P}\{\boldsymbol{Q} \in \mathcal{R}_{n,1-\alpha}\} \to 1-\alpha$.*

We now verify Assumption 2 for four standard categorical mechanisms. Each mechanism fits the same inferential pipeline; only the oracle-specific second moments change.

Example 1: direct $(k+1)$-ary randomized response (kRR). On input $c \in \{0,\dots,k\}$, direct randomized response returns the true category with probability $p = e^{\varepsilon}/(e^{\varepsilon}+k)$ and each false category with probability $q = 1/(e^{\varepsilon}+k)$. With $\rho := p-q = (e^{\varepsilon}-1)/(e^{\varepsilon}+k)$, an unbiased decoder is

$$\phi_r(y) = \frac{\mathbf{1}\{y=r\}-q}{\rho}, \qquad r = 0,\dots,k. \tag{7}$$

Example 2: optimized unary encoding (OUE). OUE first one-hot encodes the category and then perturbs each bit independently: the unique true one is reported as one with probability $p = 1/2$, while every zero bit is reported as one with probability $q = (e^{\varepsilon}+1)^{-1}$ [32]. The decoder is

$$\phi_r(y) = \frac{y_r - q}{p-q}, \qquad r = 0,\dots,k. \tag{8}$$

Example 3: optimized local hashing (OLH). OLH chooses a universal hash function H mapping $\{0,\dots,k\}$ into $\{0,\dots,g-1\}$, computes $z = H(c)$, and privatizes z via generalized randomized response on the smaller alphabet. Write

$$p_g = \frac{e^{\varepsilon}}{e^{\varepsilon}+g-1}, \qquad q_g = \frac{1}{e^{\varepsilon}+g-1}, \qquad \eta_g = q_g + \frac{p_g - q_g}{g}.$$

Treating (H, Y) as the private report, an unbiased decoder is

$$\phi_r(H,Y) = \frac{\mathbf{1}\{H(r)=Y\}-\eta_g}{p_g-\eta_g}, \qquad r = 0,\dots,k. \tag{9}$$

The optimized choice is $g \approx e^{\varepsilon}+1$, see [32].

Example 4: Hadamard response (HR). Let K be the smallest power of two strictly larger than $k+1$. Let $H_K \in \{\pm 1\}^{K\times K}$ be a Hadamard matrix and associate category $r \in \{0,\dots,k\}$ with the $(r+2)$nd row, skipping the all-ones first row. Write $h_r(\cdot)$ for the associated row function and define

$$\lambda := \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}, \qquad s := \frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}.$$

The HR report $Y \in \{1,\dots,K\}$ is drawn from

$$\mathbb{P}(Y = y \mid C = r) = \frac{1}{K}\{1 + \lambda h_r(y)\}, \tag{10}$$

which is equivalent to the usual "sample from the positive half with probability $e^{\varepsilon}/(e^{\varepsilon}+1)$ and from the negative half otherwise" construction. The decoder is $\phi_r(Y) = s h_r(Y), r = 0,\dots,k$.

Proposition 1 (Concrete oracles satisfy Assumption 2)**.** *Examples 1–4 satisfy Assumption 2. The decoder bounds can be taken as*

$$B_{\varepsilon,k}^{\mathrm{kRR}} = \frac{e^\varepsilon + k - 1}{e^\varepsilon - 1}, \qquad B_{\varepsilon,k}^{\mathrm{OUE}} = \frac{2e^\varepsilon}{e^\varepsilon - 1},$$
$$B_{\varepsilon,k}^{\mathrm{OLH}} = \frac{g(e^\varepsilon + g - 1)}{(g-1)(e^\varepsilon - 1)}, \qquad B_{\varepsilon,k}^{\mathrm{HR}} = \frac{e^\varepsilon + 1}{e^\varepsilon - 1}.$$

Moreover, the oracle-specific covariance ingredients are:

(i) ***kRR.*** *For $r \in \{0, \dots, k\}$ and $u, v \in \{0, \dots, k\}$, the conditional covariance of the decoded one-hot vector is*

$$\mathrm{Cov}\{\phi_u(Y), \phi_v(Y) \mid C = r\} = \begin{cases} \dfrac{p(1-p)}{\rho^2}, & u = v = r, \\ \dfrac{q(1-q)}{\rho^2}, & u = v \neq r, \\ -\dfrac{pq}{\rho^2}, & u = r,\ v \neq r \text{ or } v = r,\ u \neq r, \\ -\dfrac{q^2}{\rho^2}, & u \neq v,\ u \neq r,\ v \neq r. \end{cases}$$

At the truth $\boldsymbol{Q}$, let $\mu_j := jq + \rho\tau_j$. Then the total prefix covariance is

$$\Sigma_{\mathrm{kRR},jl}(\boldsymbol{Q}) = \frac{\mu_{j\wedge l}\{1 - \mu_{j\vee l}\}}{\rho^2}, \qquad j, l = 1, \dots, k. \tag{11}$$

(ii) ***OUE.*** *For $r \in \{0, \dots, k\}$, the conditional covariance of the decoded one-hot vector is diagonal:*

$$\mathrm{Var}\{\phi(Y) \mid C = r\} = \mathrm{diag}\Big(\underbrace{\frac{q(1-q)}{(p-q)^2}, \dots, \frac{q(1-q)}{(p-q)^2}}_{r \text{ entries}}, \frac{p(1-p)}{(p-q)^2}, \underbrace{\frac{q(1-q)}{(p-q)^2}, \dots, \frac{q(1-q)}{(p-q)^2}}_{k-r \text{ entries}}\Big).$$

At the truth, the total prefix covariance is

$$\Sigma_{\mathrm{OUE},jl}(\boldsymbol{Q}) = \tau_{j\wedge l}\big(1 - \tau_{j\vee l}\big) + \frac{(j \wedge l)q(1-q) + \tau_{j\wedge l}\{p(1-p) - q(1-q)\}}{(p-q)^2}, j, l = 1, \dots, k. \tag{12}$$

(iii) ***OLH.*** *For a fixed hash function H, define*

$$\pi_m^{(r,H)} := p_g \mathbf{1}\{m = H(r)\} + q_g \mathbf{1}\{m \neq H(r)\}, \qquad m = 0, \dots, g-1.$$

Then the conditional second-moment matrix is

$$[M_r^{\mathrm{OLH}}(H)]_{uv} = \frac{\mathbf{1}\{H(u) = H(v)\}\pi_{H(u)}^{(r,H)} - \eta_g \pi_{H(u)}^{(r,H)} - \eta_g \pi_{H(v)}^{(r,H)} + \eta_g^2}{(p_g - \eta_g)^2}, \tag{13}$$

so that $M_r^{\mathrm{OLH}} = \mathbb{E}_H\big[M_r^{\mathrm{OLH}}(H)\big]$, $\qquad \Gamma_r^{\mathrm{OLH}} = M_r^{\mathrm{OLH}} - \boldsymbol{e}_r \boldsymbol{e}_r^\top$.

(iv) ***HR.*** *Let h_r be the selected Hadamard rows and define*

$$[M_r^{\mathrm{HR}}]_{uv} = s^2 \frac{1}{K} \sum_{y=1}^{K} h_u(y) h_v(y)\{1 + \lambda h_r(y)\}, \tag{14}$$

so that $\Gamma_r^{\mathrm{HR}} = M_r^{\mathrm{HR}} - \boldsymbol{e}_r \boldsymbol{e}_r^\top$.

For kRR and OUE, (11) and (12) provide closed-form expressions for $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$. For OLH and HR, the displayed second-moment formulas determine $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$ once the hash family or Hadamard design is fixed.

We further discuss these examples, which differ only through the decoded prefix covariance $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$, and hence through $\Omega_{\mathcal{O}} = D^{-1}\Sigma_{\mathcal{O}}(\boldsymbol{Q})D^{-1}$ in Theorem 4. Since D is common across mechanisms for

a fixed distribution and target grid, mechanism comparison is a comparison of the oracle covariance after this common density weighting. In particular, Loewner and log-determinant comparisons may be made at the level of $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$ up to the common congruence by D^{-1}, whereas trace risk is the density-weighted quantity $\mathrm{tr}(D^{-1}\Sigma_{\mathcal{O}}(\boldsymbol{Q})D^{-1})$.

The direct kRR oracle is the closest analogue of observing a noisy categorical label. Its prefix covariance is fully explicit in (11); as $\varepsilon \to \infty$, $q \to 0$ and $\rho \to 1$, so kRR recovers the nonprivate categorical covariance $\tau_{j\wedge l}\{1 - \tau_{j\vee l}\}$. For moderate or large alphabet size, however, the factor ρ^{-2} and the false-category mass jq can substantially inflate prefix variances. Thus kRR is transparent and communication-efficient, but it is not necessarily the best oracle when many cutpoints are used.

OUE has a different structure. Conditional on the true category, the decoded coordinates are independent, which leads to the diagonal conditional covariance and the closed-form prefix covariance in (12). Its decoder bound does not grow with k, and the privacy noise accumulated by a prefix is additive over the $j \wedge l$ included bits. This often makes OUE attractive when the categorical alphabet is not very small. The price is communication: the report is a length-$k+1$ binary vector, and because the true bit is still randomized with probability $p = 1/2$, OUE does not reduce to the nonprivate oracle even when ε is large.

OLH trades the length-$k+1$ unary report for a privatized hash value on a smaller alphabet. The choice of g balances generalized randomized response noise against hash collisions; the usual optimized choice $g \approx e^{\varepsilon} + 1$ keeps the decoder bound independent of $k+1$ up to constants. Its covariance is therefore determined by two components: the randomized-response noise within the hashed alphabet and the collision structure of the chosen hash family. Formula (13) makes this dependence explicit. If the per-user hash can be generated from public or shared randomness known to the server, only the hashed response needs to be communicated; otherwise the seed or description of H is part of the report.

HR uses a fixed Hadamard design rather than random hashing. Its decoder bound is also independent of k up to constants, and the report has alphabet size K, the next power of two larger than $k+1$. The covariance formula (14) shows that HR replaces random hash collisions by deterministic algebraic correlations among selected Hadamard rows. Consequently, HR is communication-efficient and has a simple bounded decoder, but the exact prefix covariance depends on the chosen row embedding and should be evaluated through $P\Gamma_{\mathrm{HR}}P^{\top}$ for the target grid.

Overall, these comparisons do not suggest a mechanism that should be expected to dominate uniformly over all $(\varepsilon, k, \boldsymbol{\tau}, F)$. kRR is exact and approaches the nonprivate categorical oracle at large privacy budgets; OUE gives closed-form covariances with conditionally independent bit noise but uses a length-$k+1$ report; OLH reduces communication by hashing at the cost of collision-induced covariance; and HR gives a deterministic low-communication construction whose covariance is governed by Hadamard row interactions. The self-normalized confidence region has the same limiting critical value U_k for all four mechanisms. In the asymptotic theory, the mechanism affects the ellipsoid size through $\Omega_{\mathcal{O}}$; the decoder bound $B_{\varepsilon,k}$ enters the stability arguments and provides a crude finite-sample control.

4 Simulations

We report only the Polyak-Ruppert average $\bar{\boldsymbol{q}}_n$, which is the estimator covered by Theorems 4 and 5. All tables below therefore refer to the full Polyak-Ruppert average.

Setup. We instantiate Algorithm 1 with the kRR mechanism introduced in Example 1. This gives a representative local randomizer with an explicit covariance formula and lets the simulations focus on the asymptotic claims of the theory rather than on mechanism ranking. We fix $\tau = (0.1, 0.25, 0.5, 0.75, 0.9)$, and consider three data-generating distributions: $N(0,1)$, $\mathrm{Exp}(1)$, and Student-t with 3 degrees of freedom. The privacy levels are $\varepsilon \in \{1, 2, 3\}$. For each simulated distribution, we set the projection box to $[L, U] = [F^{-1}(0.001), F^{-1}(0.999)]$. We use the step-size schedule $\gamma_t = \gamma_0 t^{-0.6}$, with $\gamma_0 = 0.5(U-L)/(k+1)$, Euclidean projection onto the monotone box $\mathcal{C}$, and the box-linear initialization induced by the endpoints L and U. Every (F, ε) condition is run on the common checkpoint grid $n \in \{10^5, 3\times10^5, 10^6, 3\times10^6, 10^7, 3\times10^7\}$, with 1000 Monte Carlo paths per condition over the full grid. Oracle Gaussian coverage is a diagnostic using the known population covariance, unavailable to the feasible procedure, and is computed from Theorem 4 using

the theoretical covariance Ω_{kRR} and the $\chi^2_{5,0.95}$ critical value. Self-normalized coverage is computed from Theorem 5 using the fixed-b critical value $u^{\mathrm{SN}}_{5,0.95} \approx 359.1$.

Main findings. Table 1 shows a clean large-sample pattern. The RMSE decreases steadily and the trace-scaled risk moves toward one, from 5.714 at $n = 10^5$ to 1.046 at $n = 3 \times 10^7$. Using the four largest checkpoints $\{10^6, 3 \times 10^6, 10^7, 3 \times 10^7\}$, the average log-log RMSE slope is -0.550; for the Gaussian design the slopes are -0.507, -0.503, and -0.506 for $\varepsilon = 1, 2, 3$, respectively, which is very close to the predicted root-n rate. Projection is primarily a transient device: the mean fraction of the sample at which the final active projection occurs drops from 3.64×10^{-2} at $n = 10^5$ to 1.28×10^{-4} at $n = 3 \times 10^7$, supporting the asymptotic irrelevance result in Theorem 3.

Table 1: Pooled large-sample kRR results for the Polyak-Ruppert average. The table aggregates over all nine (F, ε) conditions, each with 1000 Monte Carlo paths. "Scaled risk" is $n \, \|\bar{\boldsymbol{q}}_n - \boldsymbol{Q}\|_2^2 / \mathrm{tr}(\Omega_{\mathrm{kRR}})$ averaged over Monte Carlo repetitions, and "last proj. frac." is the average fraction of the sample size at which the final active projection occurred.

n	RMSE	scaled risk	oracle cov.	SN cov.	last proj. frac.
10^5	0.02819	5.714	0.614	0.810	3.64×10^{-2}
3×10^5	0.01237	2.657	0.754	0.854	1.28×10^{-2}
10^6	0.00564	1.571	0.882	0.896	3.83×10^{-3}
3×10^6	0.00295	1.215	0.923	0.915	1.28×10^{-3}
10^7	0.00154	1.073	0.941	0.928	3.83×10^{-4}
3×10^7	0.000878	1.046	0.942	0.930	1.28×10^{-4}

Coverage improves with n, but the 95% target is not fully reached even at the largest checkpoint. At $n = 3 \times 10^7$, the pooled oracle Gaussian coverage is 0.942 and the pooled self-normalized coverage is 0.930. The condition-wise results at the largest checkpoint are reported in Appendix A. The hardest regime remains Student-t_3 with $\varepsilon = 1$, where oracle coverage is 0.932 and self-normalized coverage is 0.924, while moderate regimes are much closer to nominal coverage. This gradual approach to the nominal level is not unique to the many-quantile setting: related tuning-free LDP CDF inference also shows that coverage stabilizes only at fairly large sample sizes [25].

Thus the main simulation message is twofold. First, the averaged projected estimator itself is well calibrated: it exhibits near-root-n scaling and near-unit scaled risk. Second, the self-normalized region improves steadily with n but remains mildly undercovered in finite samples, indicating that calibration of the random normalizer is slower than calibration of the averaged estimator and the oracle Gaussian approximation.

Appendix A reports the condition-level results at the largest checkpoint together with an oracle-initialization diagnostic. Initializing the recursion directly at the true quantile vector does not materially change the large-sample coverage, so eliminating initial error does not resolve undercoverage in these runs; this does not establish insensitivity to arbitrary initializations or boxes.

Equal-budget comparisons. All methods in Table 2 use the same total users and per-user ε-LDP budget. Joint kRR sends one $(k+1)$-category report; composition sends k binary reports at ε/k each; sample splitting publicly assigns each user uniformly to one coordinate with one binary report at ε. At $k = 5$, their direct report encodings require three, five, and one bits, respectively, excluding common downlink queries. Trace ratios use $\Omega_{\mathcal{O}} = D^{-1}\Sigma_{\mathcal{O}}D^{-1}$ and compare asymptotic squared-error criteria. They are not finite-sample coverage results, and trace advantage does not imply dominance in every linear direction; sample splitting is not uniformly dominated in Loewner order across these settings.

Grid sensitivity. Table 3 uses Normal data, $\varepsilon = 2$, the original box, step-size rule, and six checkpoints, with 1000 paths per grid. The $k = 5$ results reuse the core experiment; $k = 3, 9$ use uniform levels $\tau_j = j/(k+1)$. Thus grid geometry and dimension both change. The critical values for $k = 3, 9$ use 50000 Brownian-functional draws and a 4000-point integration grid; the original $k = 5$ value is retained. Near-unit scaled risk and oracle coverage close to nominal coexist with material self-normalized undercoverage on the larger grid.

Table 2: Trace of asymptotic quantile covariance relative to joint kRR for standard Normal data and the core grid $(0.1, 0.25, 0.5, 0.75, 0.9)$. OUE uses Example 2; OLH uses fully random public hashing with $g = 4, 8, 21$ at $\varepsilon = 1, 2, 3$; HR uses a Sylvester matrix of order 8 with categories assigned to rows $r + 2$ in one-based indexing. These fixed designs do not give a universal mechanism ranking.

ε	composition	sample split	OUE	OLH	HR
1	7.952	1.656	3.706	3.750	5.127
2	11.618	2.801	5.162	5.100	11.627
3	12.241	3.853	5.577	5.567	19.748

Table 3: kRR grid sensitivity at $n = 3 \times 10^7$; slopes use $n \geq 10^6$. Oracle coverage uses the known asymptotic covariance. Binomial Monte Carlo standard errors are about 0.007 for oracle coverage and 0.008, 0.009, 0.011 for SN coverage.

k	RMSE slope	scaled risk	oracle cov.	SN cov.	$u_{k,0.95}$
3	-0.502	1.009	0.948	0.928	172.340
5	-0.503	1.048	0.943	0.910	359.100
9	-0.509	1.006	0.947	0.872	867.993

Appendix A.3 gives further oracle-comparison details and the salary-data point-estimation illustration.

5 Concluding remarks

We studied joint estimation and inference for a fixed finite set of quantiles under local differential privacy. By reducing the problem to interval-label estimation, we obtain a mechanism-agnostic online framework: any ε-LDP categorical oracle with an unbiased bounded one-hot decoder can be combined with projected stochastic approximation, Polyak-Ruppert averaging, and multivariate self-normalization. The procedure enforces noncrossing at every step and yields an asymptotically pivotal joint confidence region without estimating nuisance densities or asymptotic covariance matrices.

Two structural features are key. Projection onto the monotone box gives finite-sample monotonicity but is asymptotically inactive because the true quantiles lie in the interior. Efficiency comparisons reduce to the oracle-specific matrix $\Omega_{\mathcal{O}} = D^{-1}\Sigma_{\mathcal{O}}(\boldsymbol{Q})D^{-1}$, so mechanism design becomes a matrix-risk problem once decoded second moments are known.

Large-scale kRR experiments support the theory: Polyak-Ruppert averaging shows near-root-n scaling, trace-scaled risk close to one, and projection is active mainly during early transients. Oracle Gaussian coverage moves close to the nominal 95% level on the common checkpoint grid up to 3×10^7, but remains mildly low at the largest checkpoint. The self-normalized region improves with n yet remains more noticeably undercovered, suggesting slower finite-sample calibration of the random self-normalizer rather than failure of the stochastic-approximation or covariance theory. Appendix A further shows that oracle initialization does not remove undercoverage in the tested conditions, and this gradual stabilization of coverage is consistent with recent tuning-free LDP CDF inference as well [25].

These results assume fixed k and ε, distinct interior targets, and positive continuous local densities. Coverage is asymptotic and can require large samples. OUE, OLH, and HR are compared through asymptotic covariances; finite-sample coverage experiments use kRR only. Growing-k or vanishing-ε inference and finite-sample self-normalized coverage require further analysis. Other open directions include semiparametric efficiency over all ε-LDP channels, slowly increasing k with high-dimensional martingale and self-normalized limits, and general inference for non-SGD methods based on LDP CDF, range-query, or density estimators.

Acknowledgments and Disclosure of Funding

Qirui Hu's research was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 12601520), the Shanghai Engineering Research Center of Finance Intelligence (Grant No. 19DZ2254600) and by TRR 391 *Spatio-temporal Statistics for the Transition of Energy and Transport* (Project number 520388526) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). This work was also supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery Grant RGPIN-2026-06672 and Discovery Launch Supplement DGECR-2026-00288.

References

[1] Anders Aamand, Fabrizio Boninsegna, Abigail Gentle, Jacob Imola, and Rasmus Pagh. Lightweight protocols for distributed private quantile estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 27–58. PMLR, 2025.

[2] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1120–1129. PMLR, 2019.

[3] Daniel Alabi, Omri Ben-Eliezer, and Anamay Chaturvedi. Bounded space differentially private quantiles. *Transactions on Machine Learning Research*, 2023.

[4] Richard E. Barlow, David J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, 1972.

[5] Leheng Cai, Qirui Hu, and Shuyuan Wu. Federated learning of quantile inference under local differential privacy. In *International Conference on Learning Representations*, 2026.

[6] Graham Cormode, Samuel Maddock, and Carsten Maple. Frequency estimation under local differential privacy. *Proceedings of the VLDB Endowment*, 14(11):2046–2058, 2021.

[7] Bennett Cyphers and Kalyan Veeramachaneni. AnonML: Locally private machine learning over a network of peers. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 549–560. IEEE, 2017.

[8] Jörg Drechsler, Ira Globus-Harris, Audra McMillan, Jayshree Sarathy, and Adam D. Smith. Nonparametric differentially private confidence intervals for the median. *Journal of Survey Statistics and Methodology*, 10(3):804–829, 2022.

[9] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[10] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pages 371–380. ACM, 2009.

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[12] Cecilia Ferrando, Shufan Wang, and Daniel Sheldon. Parametric bootstrap for differentially private confidence intervals. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1598–1618. PMLR, 2022.

[13] Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private mean estimation: Z-test and tight confidence intervals. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2545–2554. PMLR, 2019.

[14] Jennifer Gillenwater, Matthew Joseph, and Alex Kulesza. Differentially private quantiles. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3713–3722. PMLR, 2021.

[15] Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

[16] Jacob Imola, Fabrizio Boninsegna, Hannah Keller, Anders Aamand, Amrita Roy Chowdhury, and Rasmus Pagh. Differentially private quantiles with smaller error. In *Advances in Neural Information Processing Systems*, 2025.

[17] Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth. The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 94–105, 2019.

[18] Haim Kaplan, Shachar Schnapp, and Uri Stemmer. Differentially private approximate quantiles. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10751–10761. PMLR, 2022.

[19] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[20] Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[21] Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2 edition, 2003.

[22] Clément Lalanne, Aurélien Garivier, and Rémi Gribonval. Private statistical estimation of many quantiles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18399–18418. PMLR, 2023.

[23] Clément Lalanne, Clément Gastaud, Nicolas Grislain, Aurélien Garivier, and Rémi Gribonval. Private quantiles estimation in the presence of atoms. *Information and Inference: A Journal of the IMA*, 12(3):2197–2223, 2023.

[24] Yi Liu, Qirui Hu, Lei Ding, and Linglong Kong. Online local differential private quantile inference via self-normalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21698–21714. PMLR, 2023.

[25] Yi Liu, Qirui Hu, and Linglong Kong. Tuning-free estimation and inference of cumulative distribution function under local differential privacy. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 31147–31164. PMLR, 2024.

[26] Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pages 75–84. ACM, 2007.

[27] Drago Plecko, Nicolas Bennett, and Nicolai Meinshausen. fairadapt: Causal reasoning for fair data pre-processing. *Journal of Statistical Software*, 110(4), 2024.

[28] Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[29] Herbert Robbins and David Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In J. S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.

[30] Xiaofeng Shao. A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):343–366, 2010.

[31] Lukas Steinberger. Efficiency in local differential privacy. *The Annals of Statistics*, 52(5):2139–2166, 2024.

[32] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745. USENIX Association, 2017.

[33] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

A Additional simulation results

This appendix collects the condition-level simulation tables that support Section 4. The two core simulation tables below use the same kRR oracle, target grid $\tau = (0.1, 0.25, 0.5, 0.75, 0.9)$, projection box $[L, U] = [F^{-1}(0.001), F^{-1}(0.999)]$, step-size schedule $\gamma_t = \gamma_0 t^{-0.6}$ with $\gamma_0 = 0.5(U - L)/(k + 1)$, and 95% nominal coverage level as in the main text. Every condition is run on the common checkpoint grid $n \in \{10^5, 3 \times 10^5, 10^6, 3 \times 10^6, 10^7, 3 \times 10^7\}$, with 1000 Monte Carlo paths per condition.

A.1 Condition-level results at the largest checkpoint

Table A.1: Condition-level results for the Polyak-Ruppert average at the largest checkpoint $n = 3 \times 10^7$. The Student-t_3 design with $\varepsilon = 1$ remains the slowest regime. The self-normalized coverage is below 0.95 in every condition even at the largest checkpoint, while projection is effectively inactive in all but the earliest part of the trajectory.

Distribution	ε	RMSE	scaled risk	oracle cov.	SN cov.	last proj. frac.
Exponential	1	0.00140	1.072	0.951	0.940	1.04×10^{-3}
Exponential	2	0.000553	0.983	0.938	0.939	2.60×10^{-5}
Exponential	3	0.000379	1.084	0.934	0.921	3.35×10^{-6}
Normal	1	0.00129	1.023	0.936	0.935	7.96×10^{-6}
Normal	2	0.000542	1.048	0.943	0.910	1.24×10^{-7}
Normal	3	0.000345	1.005	0.943	0.934	3.80×10^{-9}
Student-t_3	1	0.00206	1.147	0.932	0.924	7.29×10^{-5}
Student-t_3	2	0.000812	1.042	0.937	0.935	9.81×10^{-7}
Student-t_3	3	0.000521	1.010	0.960	0.935	9.05×10^{-8}

We also examine the ratio between the empirical 95% quantile of the self-normalized statistic and the fixed-b critical value. At $n = 3 \times 10^7$, this ratio remains above one in every condition and is largest for the Gaussian designs, which explains why the self-normalized coverage improves with n but does not yet match the nominal level in the main experiment.

A.2 Oracle-initialization diagnostic

To test whether the remaining undercoverage is mainly an initialization artifact, we reran five stress conditions with the same kRR configuration but initialized the algorithm directly at the true quantile vector $\boldsymbol{Q}$. This removes the initialization error by construction but leaves the stochastic approximation, projection, and privacy mechanism unchanged.

Table A.2: Box-linear versus oracle initialization at the largest checkpoint $n = 3 \times 10^7$ for five stress conditions, each with 1000 repetitions. Box and true denote box-linear and true-quantile initialization; Gaussian and SN columns report coverage. Oracle initialization drives the initial error to zero but does not materially improve the large-sample coverage.

Distribution	ε	Gaussian (box)	Gaussian (true)	SN (box)	SN (true)	ΔSN cov.
Exponential	1	0.947	0.947	0.924	0.924	0.000
Exponential	3	0.950	0.960	0.937	0.936	-0.001
Normal	2	0.936	0.935	0.941	0.940	-0.001
Student-t_3	1	0.924	0.925	0.931	0.932	$+0.001$
Student-t_3	2	0.941	0.943	0.944	0.943	-0.001

Pooled over these five conditions at $n = 3 \times 10^7$, oracle coverage changes only from 0.940 to 0.942, while self-normalized coverage changes from 0.935 to 0.935. Thus removing initial error does not resolve undercoverage in these conditions. This diagnostic does not identify a unique cause or establish robustness to arbitrary initialization.

A.3 Additional comparisons and empirical results

The equal-budget comparison and grid-sensitivity results are reported in Tables 2 and 3 in Section 4.

The oracle calculations use Proposition 1; exact finite-alphabet enumeration (equivalent hash-marginal sums for OLH) checks decoder means and second moments. OLH communication assumes a public/shared hash; transmitting a seed adds cost. At small ε, kRR has $\rho \sim \varepsilon/(k+1)$, suggesting coordinate error of scale $k/(\varepsilon\sqrt{n})$ up to density and grid factors. For fixed k and vanishing ε_n, the analogous scaling suggests $n\varepsilon_n^2 \to \infty$, but this is not a proved triangular-array limit. Under stronger global density conditions, projected-SA bounds offer a route to finite-sample estimation guarantees [21]; coverage of the random self-normalizer requires additional analysis.

Salary-data point estimation. We use `gov_census` from `fairadapt` 1.0.0 [27], distributed by Plecko and Bennett under the package's GPL ($\geq$ 3) license. The package documentation describes government-employee salaries from the 2018 US Census. Retaining annual salaries at most USD 200000 leaves 202958 records with 737 distinct values; no jitter is added. The full filtered-data inverse-ECDF reference quantiles at the core five levels are USD 14200, 30000, 49000, 71000, and 100000, used only for evaluation.

Each of 100 independent repetitions uses a random permutation without replacement, shared across methods, and fresh private reports. The primary comparison uses its first 200000 records. All private methods use box $[0, 200000]$, box-linear initialization, the original step-size rule, isotonic projection, and full averaging without method-specific retuning. The sample-split method multiplies its selected stochastic score by k. All methods use the same total users and per-user privacy budget.

Table A.3: Salary-data RMSE in thousands of dollars at $n = 200000$, computed as the square root of mean squared error over all five empirical reference targets and 100 repetitions.

Method	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$
Joint kRR	0.77	0.31	0.24
Sample split	1.07	0.51	0.38
Composition	2.46	1.12	0.68

Joint kRR RMSE decreases from $11.83/6.69/6.48$ at $n = 5000$ to $0.77/0.31/0.24$ at $n = 200000$ for $\varepsilon = 1/2/3$, and every joint estimate is noncrossing. This is a finite-data point-estimation illustration. The ties and sampling without replacement do not satisfy the stated i.i.d. smooth-density inference assumptions and do not validate confidence-region coverage.

B Proofs

This appendix contains formal proofs of all theorems and propositions stated in the main text.

B.1 Auxiliary lemmas

Lemma 1. *Let $\mathcal{C}$ be a nonempty closed convex subset of $\mathbb{R}^k$, and let $\Pi_{\mathcal{C}}$ denote Euclidean projection onto $\mathcal{C}$. Then for all $x, y \in \mathbb{R}^k$,*

$$\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(y)\| \leq \|x - y\|.$$

Moreover, for every $x \in \mathbb{R}^k$ and every $z \in \mathcal{C}$,

$$\langle x - \Pi_{\mathcal{C}}(x), z - \Pi_{\mathcal{C}}(x)\rangle \leq 0.$$

Proof. Both statements are standard. The second is the first-order optimality condition for Euclidean projection onto a closed convex set. Applying it twice with $(x, z) = (x, \Pi_{\mathcal{C}}(y))$ and $(y, z) = (y, \Pi_{\mathcal{C}}(x))$ and adding the two inequalities yields the first. □

Lemma 2. *Under Assumption 2,*

$$\left\|\widehat{G}_t\right\|_\infty \leq kB_{\varepsilon,k} \quad \textit{almost surely for every } t.$$

Consequently,

$$\left\|\boldsymbol{\tau} - \widehat{G}_t\right\|_\infty \le 1 + kB_{\varepsilon,k} \quad \textit{almost surely.}$$

Proof. The jth coordinate of $\widehat{G}_t = P\phi(Y_t)$ is a sum of at most $j \le k$ coordinates of $\phi(Y_t)$. Therefore

$$|\widehat{G}_{j,t}| \le j \left\|\phi(Y_t)\right\|_\infty \le kB_{\varepsilon,k}$$

a.s. for each j. □

B.2 Proof of Proposition 1

Proof of Proposition 1. We verify each mechanism separately.

(i) kRR. The privacy guarantee is immediate because for any two inputs the likelihood ratio between any two outputs is at most $p/q = e^\varepsilon$. The decoder (7) is unbiased because, if $Y \sim \mathcal{R}(r)$,

$$\mathbb{E}[\mathbf{1}\{Y = u\}] = q + \rho\mathbf{1}\{u = r\},$$

so

$$\mathbb{E}[\phi_u(Y)] = \frac{q + \rho\mathbf{1}\{u = r\} - q}{\rho} = \mathbf{1}\{u = r\}.$$

The bound follows from

$$\max\Big\{\frac{1-q}{\rho}, \frac{q}{\rho}\Big\} = \frac{e^\varepsilon + k - 1}{e^\varepsilon - 1}.$$

Since $\phi_u(Y)$ differs from the indicator $\mathbf{1}\{Y = u\}$ only by an additive constant and a scalar factor, its covariance is the covariance of the indicator vector divided by ρ^2. This gives the stated Γ_r^{kRR}. At the truth, the decoded prefix variable equals

$$\widehat{G}_j = \frac{Z_j - jq}{\rho}, \qquad Z_j := \mathbf{1}\{Y \le j - 1\},$$

where $\mathbb{E}[Z_j] = \mu_j = jq + \rho\tau_j$ and $\mathbb{E}[\widehat{G}_j] = \tau_j$. Since $Z_j \le Z_l$ for $j \le l$,

$$\mathrm{Cov}(Z_j, Z_l) = \mu_j(1 - \mu_l),$$

which yields (11).

(ii) OUE. For one-hot inputs, the OUE mechanism is ε-LDP. Indeed, if the true categories are $u \ne v$, then under the two inputs all output bits except the uth and vth have the same distribution, so the global likelihood ratio reduces to the product of the likelihood ratios of these two bits. With $p = 1/2$ and $q = (e^\varepsilon + 1)^{-1}$, the largest ratio is

$$\frac{p(1-q)}{q(1-p)} = e^\varepsilon,$$

which proves ε-LDP; see also [32]. The decoder (8) is unbiased by construction. Since $y_r \in \{0, 1\}$,

$$|\phi_r(y)| \le \max\Big\{\frac{q}{p-q}, \frac{1-q}{p-q}\Big\} = \frac{2e^\varepsilon}{e^\varepsilon - 1}.$$

Conditional on the true category r, the bits are independent, the rth bit has variance $p(1-p)$, and all others have variance $q(1-q)$. This gives the diagonal form of Γ_r^{OUE}. Because the conditional covariance is diagonal,

$$\sum_{u=0}^{j\wedge l-1} \mathbb{E}\big[\mathrm{Var}\{\phi_u(Y) \mid C\}\big] = \sum_{u=0}^{j\wedge l-1} \Big\{q(1-q) + \pi_u(\boldsymbol{Q})\big[p(1-p) - q(1-q)\big]\Big\}(p-q)^{-2}.$$

The unconditional covariance contributed by the random category itself is

$$\mathrm{Cov}\big(\mathbf{1}\{C \le j-1\}, \mathbf{1}\{C \le l-1\}\big) = \tau_{j\wedge l}\big(1 - \tau_{j\vee l}\big).$$

Since $\sum_{u=0}^{j\wedge l-1} \pi_u(\boldsymbol{Q}) = \tau_{j\wedge l}$, combining the two pieces yields (12).

(iii) OLH. The privacy guarantee follows because conditional on the hash value $H(c)$, the report is produced by generalized randomized response on a range-g alphabet. Unbiasedness of (9) is the standard OLH identity from [32]. The uniform bound follows from

$$|\phi_r(H, Y)| \le (p_g - \eta_g)^{-1} = \frac{g(e^\varepsilon + g - 1)}{(g-1)(e^\varepsilon - 1)}.$$

The second-moment formula (13) is obtained by expanding

$$\mathbb{E}\big[(\mathbf{1}\{H(u) = Y\} - \eta_g)(\mathbf{1}\{H(v) = Y\} - \eta_g) \mid C = r, H\big],$$

using

$$\mathbb{E}[\mathbf{1}\{H(u) = Y\} \mid C = r, H] = \pi^{(r,H)}_{H(u)}$$

and

$$\mathbb{E}[\mathbf{1}\{H(u) = Y\}\mathbf{1}\{H(v) = Y\} \mid C = r, H] = \mathbf{1}\{H(u) = H(v)\}\pi^{(r,H)}_{H(u)}.$$

Subtracting $\boldsymbol{e}_r \boldsymbol{e}_r^\top$ yields Γ_r^{OLH}.

(iv) HR. Equation (10) implies

$$\frac{\mathbb{P}(Y = y \mid C = r)}{\mathbb{P}(Y = y \mid C = r')} \le \frac{1 + \lambda}{1 - \lambda} = e^\varepsilon,$$

so HR is ε-LDP. Unbiasedness follows from orthogonality of Hadamard rows:

$$\mathbb{E}[\phi_u(Y) \mid C = r] = s\frac{1}{K}\sum_{y=1}^{K} h_u(y)\{1 + \lambda h_r(y)\} = s\lambda\mathbf{1}\{u = r\} = \mathbf{1}\{u = r\}.$$

The decoder bound is immediate because $|h_u(y)| = 1$. The second-moment formula (14) follows by direct substitution. □

B.3 Proof of Theorem 1

Proof of Theorem 1. Part (i): conditional on $\boldsymbol{q}_{t-1}$, the mapping $X_t \mapsto C_t(\boldsymbol{q}_{t-1})$ is deterministic. Hence the local privacy of $X_t \mapsto Y_t$ is inherited from the local privacy of the categorical randomizer $\mathcal{R}$ in Assumption 2(i).

Part (ii): conditional on $\boldsymbol{q}_{t-1}$,

$$\mathbb{E}[\widehat{v}_t \mid \boldsymbol{q}_{t-1}] = \sum_{r=0}^{k} \pi_r(\boldsymbol{q}_{t-1})\boldsymbol{e}_r = \pi(\boldsymbol{q}_{t-1}).$$

Applying the fixed matrix P gives

$$\mathbb{E}[\widehat{G}_t \mid \boldsymbol{q}_{t-1}] = P\pi(\boldsymbol{q}_{t-1}) = G(\boldsymbol{q}_{t-1}).$$

Part (iii): by construction, (3) is the projection of a point in $\mathbb{R}^k$ onto $\mathcal{C}$, so $\boldsymbol{q}_t \in \mathcal{C}$ for every t. Since $\mathcal{C}$ is convex, the average of points in $\mathcal{C}$ also lies in $\mathcal{C}$. □

B.4 Proof of Theorem 2

Proof of Theorem 2. Define the martingale-difference noise

$$M_t := G(\boldsymbol{q}_{t-1}) - \widehat{G}_t, \qquad \mathbb{E}[M_t \mid \boldsymbol{q}_{t-1}] = 0,$$

where the conditional mean-zero property follows from Theorem 1(ii). Let

$$V_t := \|\boldsymbol{q}_t - \boldsymbol{Q}\|^2.$$

Using Lemma 1 with $z = \boldsymbol{Q} \in \mathcal{C}$ and the update (3),

$$V_t \le \|\boldsymbol{q}_{t-1} - \boldsymbol{Q} + \gamma_t\{h(\boldsymbol{q}_{t-1}) + M_t\}\|^2.$$

Expanding and taking the conditional expectation gives

$$\mathbb{E}[V_t \mid \boldsymbol{q}_{t-1}] \le V_{t-1} + 2\gamma_t \langle \boldsymbol{q}_{t-1} - \boldsymbol{Q}, h(\boldsymbol{q}_{t-1}) \rangle + \gamma_t^2 K_0, \tag{15}$$

where $K_0 < \infty$ because $\boldsymbol{q}_{t-1} \in \mathcal{C} \subset [L, U]^k$ and Lemma 2 bounds M_t uniformly.

For each coordinate,

$$(q_j - Q_j)\{\tau_j - F(q_j)\} \le 0,$$

because the scalar map $q \mapsto \tau_j - F(q)$ is nonincreasing and vanishes at Q_j. Summing over coordinates yields

$$\langle \boldsymbol{q} - \boldsymbol{Q}, h(\boldsymbol{q}) \rangle \le 0 \qquad \text{for every } \boldsymbol{q} \in \mathcal{C}, \tag{16}$$

with equality if and only if $\boldsymbol{q} = \boldsymbol{Q}$.

Combining (15) and (16),

$$\mathbb{E}[V_t \mid \boldsymbol{q}_{t-1}] \le V_{t-1} + K_0 \gamma_t^2.$$

Because $\sum_t \gamma_t^2 < \infty$ by Assumption 3, the Robbins–Siegmund theorem [29] implies that V_t converges almost surely to a finite random variable and that

$$\sum_{t=1}^{\infty} -\gamma_t \langle \boldsymbol{q}_{t-1} - \boldsymbol{Q}, h(\boldsymbol{q}_{t-1}) \rangle < \infty \qquad \text{a.s.}$$

Because $\sum_t \gamma_t = \infty$, the preceding summability forces $V_t \to 0$. Indeed, suppose that $V_t \to V_\infty > 0$ on an event of positive probability. Then there exists $\eta > 0$ such that $\|\boldsymbol{q}_t - \boldsymbol{Q}\| \ge \eta$ eventually on this event. By monotonicity of F and the uniqueness of each root $F(q) = \tau_j$ in $[L, U]$, the continuous-distance-away drift satisfies

$$\inf_{\boldsymbol{q} \in \mathcal{C}:\, \|\boldsymbol{q} - \boldsymbol{Q}\| \ge \eta} \{- \langle \boldsymbol{q} - \boldsymbol{Q}, h(\boldsymbol{q}) \rangle\} > 0.$$

This contradicts

$$\sum_{t=1}^{\infty} -\gamma_t \langle \boldsymbol{q}_{t-1} - \boldsymbol{Q}, h(\boldsymbol{q}_{t-1}) \rangle < \infty$$

and $\sum_t \gamma_t = \infty$. Hence $V_t \to 0$ almost surely, i.e., $\boldsymbol{q}_t \to \boldsymbol{Q}$. The convergence $\bar{\boldsymbol{q}}_t \to \boldsymbol{Q}$ follows by Cesaro averaging. □

B.5 Proof of Theorem 3

Proof of Theorem 3. Let

$$\delta_0 := \frac{1}{4} \min\{Q_1 - L,\ Q_2 - Q_1, \ldots, Q_k - Q_{k-1},\ U - Q_k\} > 0.$$

By Theorem 2, there exists an almost surely finite random time T_1 such that for all $t \ge T_1$,

$$\|\boldsymbol{q}_{t-1} - \boldsymbol{Q}\|_\infty \le \delta_0.$$

Hence $\boldsymbol{q}_{t-1}$ lies at least $2\delta_0$ away from every boundary face of $\mathcal{C}$.

By Lemma 2,

$$\left\| \boldsymbol{\tau} - \widehat{G}_t \right\|_\infty \le 1 + k B_{\varepsilon,k} \qquad \text{a.s.}$$

Since $\gamma_t \to 0$, there exists a deterministic T_2 such that for all $t \ge T_2$,

$$\gamma_t (1 + k B_{\varepsilon,k}) \le \delta_0.$$

Therefore, for $t \ge T := \max(T_1, T_2)$, the pre-projection point

$$\widetilde{q}_t := \boldsymbol{q}_{t-1} + \gamma_t (\boldsymbol{\tau} - \widehat{G}_t)$$

remains in $\mathcal{C}$: each coordinate stays within $[L, U]$, and the difference between adjacent coordinates remains nonnegative because the perturbation to each coordinate is at most δ_0 in absolute value, while the current gap is at least $2\delta_0$. Hence $\Pi_{\mathcal{C}}(\widetilde{q}_t) = \widetilde{q}_t$ for all $t \ge T$. □

B.6 Proof of Theorem 4

Proof of Theorem 4. By Theorem 3, after an almost surely finite random time the recursion is unprojected:

$$\boldsymbol{q}_t = \boldsymbol{q}_{t-1} + \gamma_t\{h(\boldsymbol{q}_{t-1}) + M_t\}.$$

Define $\Delta_t := \boldsymbol{q}_t - \boldsymbol{Q}$. Because $h(\boldsymbol{Q}) = 0$ and F is continuously differentiable near $\boldsymbol{Q}$,

$$h(\boldsymbol{q}) = -D(\boldsymbol{q} - \boldsymbol{Q}) + r(\boldsymbol{q}), \qquad \|r(\boldsymbol{q})\| = o\big(\|\boldsymbol{q} - \boldsymbol{Q}\|\big) \quad \text{as } \boldsymbol{q} \to \boldsymbol{Q}.$$

Thus the recursion becomes

$$\Delta_t = (I - \gamma_t D)\Delta_{t-1} + \gamma_t M_t + \gamma_t r(\boldsymbol{q}_{t-1}). \tag{17}$$

The noise M_t is a bounded martingale difference sequence. Indeed, Theorem 1(ii) gives the conditional mean-zero property and Lemma 2 provides a uniform bound on the increments. Since $\boldsymbol{q}_t \to \boldsymbol{Q}$ almost surely, the conditional covariance of M_t converges almost surely to

$$\Sigma_{\mathcal{O}}(\boldsymbol{Q}) = P\Big\{\sum_{r=0}^{k} \pi_r(\boldsymbol{Q}) M_r^{\mathcal{O}} - \pi(\boldsymbol{Q})\pi(\boldsymbol{Q})^\top\Big\}P^\top.$$

The matrix D is positive definite by Assumption 1. Therefore the standard Polyak theorem for averaged stochastic approximation applies; see [28, Theorem 1]. Verifying its conditions with (17) yields

$$\sqrt{n}(\bar{\boldsymbol{q}}_n - \boldsymbol{Q}) = D^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} M_t + o_p(1).$$

Since M_t is bounded and its conditional covariance converges almost surely to $\Sigma_{\mathcal{O}}(\boldsymbol{Q})$, the multivariate martingale central limit theorem [15] gives

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} M_t \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathcal{O}}(\boldsymbol{Q})).$$

Slutsky's theorem yields the stated CLT. □

B.7 Proof of Corollary 1

Proof of Corollary 1. Theorem 4 shows that the asymptotic distribution of $\bar{\boldsymbol{q}}_n$ under oracle $\mathcal{O}$ is $\mathcal{N}(0, \Omega_{\mathcal{O}})$ after $\sqrt{n}$ scaling. Therefore any monotone matrix-risk functional Ψ is minimized asymptotically by minimizing $\Psi(\Omega_{\mathcal{O}})$ over the admissible oracle class. □

B.8 Proof of Theorem 5

Proof of Theorem 5. By Theorem 4 and the corresponding functional martingale central limit theorem for the averaged stochastic approximation process [15, 21, 28],

$$\mathbb{S}_n(r) := \frac{\lfloor nr \rfloor}{\sqrt{n}}\big(\bar{\boldsymbol{q}}_{\lfloor nr \rfloor} - \boldsymbol{Q}\big), \qquad r \in [0, 1],$$

converges weakly in $\ell^\infty([0,1]; \mathbb{R}^k)$ to $\Omega_{\mathcal{O}}^{1/2}\mathbf{B}(r)$. In particular,

$$\sqrt{n}(\bar{\boldsymbol{q}}_n - \boldsymbol{Q}) = \mathbb{S}_n(1) + o_p(1) \xrightarrow{d} \Omega_{\mathcal{O}}^{1/2}\mathbf{B}(1).$$

The self-normalizer satisfies the Riemann-sum representation

$$W_n = \frac{1}{n}\sum_{t=1}^{n}\left\{\frac{t}{\sqrt{n}}(\bar{\boldsymbol{q}}_t - \boldsymbol{Q}) - \frac{t}{n}\sqrt{n}(\bar{\boldsymbol{q}}_n - \boldsymbol{Q})\right\}\left\{\frac{t}{\sqrt{n}}(\bar{\boldsymbol{q}}_t - \boldsymbol{Q}) - \frac{t}{n}\sqrt{n}(\bar{\boldsymbol{q}}_n - \boldsymbol{Q})\right\}^\top + o_p(1).$$

By the continuous mapping theorem,

$$W_n \xrightarrow{d} \Omega_{\mathcal{O}}^{1/2} V \Omega_{\mathcal{O}}^{1/2},$$

where V is defined in (6). Since V is almost surely positive definite, so is the limit of W_n. Therefore

$$W_n^{-1} \xrightarrow{d} \Omega_{\mathcal{O}}^{-1/2} V^{-1} \Omega_{\mathcal{O}}^{-1/2}.$$

Combining the limits for $\sqrt{n}(\bar{\boldsymbol{q}}_n - \boldsymbol{Q})$ and W_n^{-1} yields

$$T_n(\boldsymbol{Q}) \xrightarrow{d} \mathbf{B}(1)^\top V^{-1}\mathbf{B}(1) = U_k.$$

The coverage statement follows because $u_{k,1-\alpha}$ is the $(1-\alpha)$ quantile of U_k. □

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: The abstract and introduction state the proposed online multi-quantile LDP framework, the noncrossing projected recursion, the asymptotic inference guarantees, the four concrete oracle examples, and the numerical evaluation. These claims are developed in Sections 2–4 and supported by Appendix B.

   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: The manuscript states the fixed-k, distributional, oracle, and step-size assumptions in Assumptions 1–3. It also discusses finite-sample undercoverage, the slower calibration of the self-normalizer, and open directions such as efficiency, increasing k, and non-SGD LDP methods in Sections 4 and 5.

   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: The theoretical results are stated under explicit assumptions in Section 3, and formal proofs are provided in Appendix B. The appendix proves the oracle examples, privacy, consistency, asymptotic inactivity of projection, the joint CLT, and the self-normalized limit.

   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: [Yes]

   Justification: Section 4 specifies the mechanism, target grid, data-generating distributions, privacy levels, step-size schedule, initialization, checkpoints, number of Monte Carlo repetitions, and critical values. Appendix A gives condition-level and robustness tables; synthetic data are generated from fully specified distributions, and Appendix A.3 describes the additional comparisons and salary-data protocol.

   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: [Yes]

   Justification: The supplementary material includes a reproduction code package.

   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: [Yes]

   Justification: The experimental setting is described in Section 4, including the kRR mechanism, $\tau = (0.1, 0.25, 0.5, 0.75, 0.9)$, distributions, $\varepsilon \in \{1, 2, 3\}$, step sizes, projection set, initialization, checkpoints, repetitions, and calibration rules. Appendix A provides additional condition-level details.

   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: [Yes]

   Justification: The tables report Monte Carlo repetitions and summary quantities such as RMSE, scaled risk, and coverage.

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: [Yes]

   Justification: The manuscript specifies sample sizes and Monte Carlo repetitions. The simulation code is lightweight and uses standard scientific Python libraries.

   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 combines theory, synthetic simulations, and secondary analysis of public salary data; it does not recruit participants or scrape personal records and is aimed at privacy-preserving inference. No aspect of the current manuscript appears to require a deviation from the NeurIPS Code of Ethics.

   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: [Yes]

    Justification: The manuscript motivates positive societal impact through privacy-preserving quantile inference under LDP.

    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 paper does not release high-risk models, scraped datasets, pretrained generative models, or other assets with a high risk of misuse. The empirical component uses synthetic distributions and public salary data distributed with `fairadapt`.

    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: [Yes]

Justification: The salary illustration credits `fairadapt` 1.0.0 and its authors [27]; Appendix A.3 specifies the source and GPL ($\geq$ 3) package license. Prior mechanisms are credited 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: [Yes]

Justification: The supplementary material includes a reproduction code package with a README, dependency file, simulation scripts, and analysis scripts. The package documents how to run the main experiment and the oracle-initialization diagnostic using synthetic data generated from the specified distributions.

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 paper does not involve crowdsourcing, user studies, surveys, or research with human subjects. Therefore participant instructions, screenshots, and compensation details are not applicable.

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 paper does not involve human-subject research or crowdsourcing. Therefore IRB approval or equivalent review 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: [N/A]

    Justification: The core methodology, theory, and experiments do not use LLMs as an important, original, or non-standard component. Any ordinary writing, editing, or formatting assistance is outside the declaration requirement described in the checklist question.

    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.