SUPPORT VECTOR MACHINES WITH RANDOMIZED RESPONSE

Anonymous authors
Paper under double-blind review

ABSTRACT

The linear support vector machine (SVM) is a fundamental binary classification model. In many classification applications, labels are sensitive, and randomized response can ensure label differential privacy (LabelDP). This paper studies statistical inference for the linear SVM under randomized response. We construct a corrected loss to estimate the linear SVM parameter. Since the hinge loss is non-differentiable, we apply convolution smoothing to obtain a differentiable objective and the corresponding estimator. For the estimator, we analyze its asymptotic properties and establish confidence intervals. We then select the randomized response probabilities by minimizing the trace of the covariance contribution induced by randomized response within the feasible regions of ε-LabelDP and (ε, δ)-LabelDP, leading to optimal randomized response mechanisms. Simulation studies and a real-data application show that the proposed method improves parameter estimation, confidence interval coverage, and prediction accuracy.

1 INTRODUCTION

The support vector machine (SVM) is one of the most fundamental binary classification models. Since the seminal work of Cortes & Vapnik (1995), the linear SVM has been widely applied to biomedicine and finance, such as cancer classification from microarray data (Guyon et al., 2002), heart failure prediction (Ali et al., 2019), major depression identification from functional neuroimaging data (Zeng et al., 2012) and credit default prediction (Huang et al., 2007). With its widespread application, the linear SVM has become increasingly important.

In many practical problems, the covariates are not sensitive and can be shared or stored openly, whereas the response label may be clinically or financially confidential information. Consequently, the true labels cannot be released directly. A possible way to protect such labels is randomized response (Warner, 1965), which replaces the true binary label by a randomized label. The respondent reports the true label with a prespecified probability and the opposite label otherwise. As a result, the analyst observes only a private label (Blair et al., 2015). Specifically, randomized response specifies how the label is perturbed, whereas label differential privacy (LabelDP) specifies the privacy constraints that the randomized response probabilities should satisfy in order to provide rigorous privacy protection (Ghazi et al., 2021; Busa-Fekete et al., 2023; Zhang et al., 2025).

Several works have studied differential privacy via randomized response. Holohan et al. (2017) studied randomized response under differential privacy constraints. Ghazi et al. (2021) proposed randomized response with prior for deep learning under LabelDP. Tang et al. (2022) developed a framework for denoising private labels before downstream training. Xu et al. (2023) analyzed binary classification under local LabelDP with randomized response and established generalization bounds and consistency guarantees. These studies have mainly focused on the design of randomized response mechanisms and machine learning with privatized labels. Zhang et al. (2025) further developed estimation and inference for generalized linear models under LabelDP.

The linear SVM is an important classification model, and usually its labels require privacy protection. Therefore, our first goal is to choose the randomized response probabilities within the feasible regions of ε-LabelDP and (ε, δ)-LabelDP. Additionally, when labels are privatized via randomized response within LabelDP framework, our goal is to estimate the linear SVM parameter, establish the asymptotic properties of the parameter estimator, and construct asymptotically valid confidence intervals. Our main contributions are as follows.

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

1. **Estimation with privatized labels**. Based on privatized labels, we construct a corrected loss whose population minimizer equals the linear SVM parameter. Since the hinge loss is not differentiable, we apply convolution smoothing (Rybak et al., 2025) to the corrected loss to obtain a differentiable objective. We establish local strong convexity in a neighborhood of the target parameter with probability tending to one, which ensures the existence of a unique local estimator in this neighborhood.
2. **Asymptotic inference under randomized response**. We establish the asymptotic normality of the estimator under randomized response and derive its asymptotic covariance matrix. We further construct a consistent estimator of the asymptotic covariance matrix. In particular, asymptotically valid confidence intervals are obtained for individual components of the linear SVM parameter.
3. **Optimal randomized response mechanisms under LabelDP**. We analyze the asymptotic covariance of the parameter estimator and isolate the covariance contribution induced by randomized response from it. We further derive optimal randomized response mechanisms under both ε-LabelDP and (ε, δ)-LabelDP by minimizing the trace of the covariance contribution induced by randomized response. This directly links mechanism design to statistical inference.

The remainder of the paper is organized as follows. Section 2 develops the SVM within the randomized response framework. Section 3 establishes privacy preserving inference. Section 4 studies optimal randomized response mechanisms under both ε-LabelDP and (ε, δ)-LabelDP constraints. Section 5 presents simulation and real-data results. Section 6 concludes the paper.

2 Linear SVM with Randomized Response

In this section, we develop the estimation framework for the linear SVM when the label is privatized via the randomized response mechanism.

Let $X \in \mathbb{R}^{d+1}$ denote the augmented feature vector, whose first coordinate is fixed at 1, let $Y^* \in \{-1, +1\}$ denote the true binary label, and let $l(u) = (u)_+$. The linear SVM parameter is defined as

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d+1}} \mathbb{E}\{l(1 - Y^* X^\top \theta)\},$$

and $\theta = (\theta_0, \theta_1^\top)^\top$, $\theta_1^\top = (\theta_{11}, \ldots, \theta_{1d}) \in \mathbb{R}^d$, $\theta_0 \in \mathbb{R}$.

The randomized response mechanism is a conditional probability rule that maps the true label Y^* to a privatized label $Y \in \{-1, +1\}$. The mechanism is characterized by two probabilities

$$p_{-1} := \Pr(Y = -1 \mid Y^* = -1), \qquad p_{+1} := \Pr(Y = +1 \mid Y^* = +1).$$

Equivalently, $\Pr(Y = -1 \mid Y^* = +1) = 1 - p_{+1}$, $\Pr(Y = +1 \mid Y^* = -1) = 1 - p_{-1}$.

To estimate θ^* with the privatized label Y, we construct a corrected loss function $\tilde{l}(\theta; X, Y)$ satisfying

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d+1}} \mathbb{E}\{l(1 - Y^* X^\top \theta)\} = \arg\min_{\theta \in \mathbb{R}^{d+1}} \mathbb{E}\{\tilde{l}(\theta; X, Y)\}.$$

A sufficient condition is

$$\mathbb{E}\{\tilde{l}(\theta; X, Y) \mid X, Y^*\} = l(1 - Y^* X^\top \theta). \tag{1}$$

Calculating (1) for $Y^* = +1$ and $Y^* = -1$ separately yields

$$\begin{pmatrix} p_{+1} & 1 - p_{+1} \\ 1 - p_{-1} & p_{-1} \end{pmatrix} \begin{pmatrix} \tilde{l}(\theta; X, +1) \\ \tilde{l}(\theta; X, -1) \end{pmatrix} = \begin{pmatrix} l(1 - X^\top \theta) \\ l(1 + X^\top \theta) \end{pmatrix}.$$

This leads to the corrected loss function

$$\tilde{l}(\theta; X, Y) = \begin{cases} \dfrac{p_{-1}\, l(1 - X^\top \theta) - (1 - p_{+1})\, l(1 + X^\top \theta)}{p_{+1} + p_{-1} - 1}, & Y = +1, \\ \dfrac{p_{+1}\, l(1 + X^\top \theta) - (1 - p_{-1})\, l(1 - X^\top \theta)}{p_{+1} + p_{-1} - 1}, & Y = -1. \end{cases} \tag{2}$$

Motivated by Rybak et al. (2025), we apply convolution smoothing to hinge loss $l(u)$. Let $K(\cdot) : \mathbb{R} \to [0, \infty)$ be a symmetric kernel density function and $h > 0$ be a bandwidth parameter. The convolution-smoothed hinge loss, denoted as $l_h(u)$, is defined as

$$l_h(u) = \int_{-\infty}^{\infty} l(t) \frac{1}{h} K\left(\frac{u-t}{h}\right) dt.$$

In (2), replacing $l(\cdot)$ with $l_h(\cdot)$ gives $\tilde{l}_h(\theta; X, Y)$. For the available dataset $\{X_i, Y_i\}_{i=1}^n$, we define the empirical objective $M_n(\theta)$ and the estimator $\hat{\theta}_h$,

$$\hat{\theta}_h = \underset{\theta \in \mathbb{R}^{d+1}}{\arg\min}\, M_n(\theta) = \underset{\theta \in \mathbb{R}^{d+1}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \tilde{l}_h(\theta; X_i, Y_i) + \frac{\lambda}{2} \theta^\top J \theta \right\}. \tag{3}$$

where $\lambda > 0$ is a regularization parameter, $J = \mathrm{diag}(0, I_d)$.

3 Theoretical Properties

In this section, we establish the asymptotic theory for the linear SVM with randomized response. Throughout the paper, for any vector v and any admissible index s, v_{-s} denotes the subvector obtained by deleting the coordinate indexed by s, and $(v)_s$ denotes its s-th component. We next introduce the following assumptions.

Assumption 1. *The randomized response mechanism satisfies $p_{-1} + p_{+1} - 1 > 0$.*

Assumption 2. *The kernel function $K(\cdot)$ satisfies the following conditions:*

(i) $K(\cdot) \geq 0$, $\int_{\mathbb{R}} K(u)du = 1$, and $K(u) = K(-u)$;

(ii) $\|K(\cdot)\|_\infty < \infty$, $K(\cdot) \in C^1(\mathbb{R})$, $\|K'(\cdot)\|_\infty < \infty$, $\int |u| K(u)du < \infty$ and $\int u^2 K(u)du < \infty$;

(iii) There exist constants $\varepsilon_K > 0$ and $c_K > 0$ such that $\inf_{|u| \leq \varepsilon_K} K(u) \geq c_K > 0$.

Assumption 3. *The minimizer $\theta^* = (\theta_0^*, (\theta_1^*)^\top)^\top$ of the linear SVM population loss function is unique, where $\theta_1^* = (\theta_{11}^*, \ldots, \theta_{1d}^*)^\top \in \mathbb{R}^d$. For some $s \in \{1, \ldots, d\}$, $\theta_{1s}^* \neq 0$.*

Assumption 4. *The conditional random vector $X_{-1} \mid Y^* = y$ has a density $f_y(\cdot)$ on $\mathbb{R}^d$ and for some constant $r_0 > 0$, define $\mathcal{N} := B(\theta^*, r_0)$. For $j = 0, 1$, assume that $\partial^j_{(x_{-1})_s} f_y(x_{-1})$ exists and is continuous, and there exists an integrable envelope function $M_{y,j}((x_{-1})_{-s})$ satisfying: for all $\theta \in \mathcal{N}$, all $u \in \mathbb{R}$, and almost all $(x_{-1})_{-s} \in \mathbb{R}^{d-1}$, it holds that*

$$\left(1 + \|x((x_{-1})_{-s}, \chi_y)\|_2^2\right) \left|\partial^j_{(x_{-1})_s} f_y((x_{-1})_{-s}, \chi_y)\right| \leq M_{y,j}((x_{-1})_{-s})$$

where $\chi_y = \chi_y((x_{-1})_{-s}, u; \theta) := \dfrac{1 - y\theta_0 - y(\theta_1)_{-s}{}^\top (x_{-1})_{-s} - u}{y\theta_{1s}}$, and $x((x_{-1})_{-s}, (x_{-1})_s) \in \mathbb{R}^{d+1}$ denotes the augmented vector obtained by restoring the intercept term 1 and inserting $(x_{-1})_s$ back into the s-th coordinate.

Assumption 5. *Let $f_+(\cdot) = f_y(\cdot)$ for $y = +1$, $f_-(\cdot) = f_y(\cdot)$ for $y = -1$ and $A_s \in \mathbb{R}^{d \times d}$ be an orthogonal matrix such that $A_s \theta_1^* / \|\theta_1^*\|_2 = e_s$, where e_s is the s-th standard basis vector. Let $z = A_s x_{-1}$. Suppose there exists a rectangle $R = \prod_{j \neq s} [l_j, v_j] \subset \mathbb{R}^{d-1}$ and constants $c_+, c_- > 0$, such that*

$$f_+(A_s^\top z) \geq c_+, \qquad \textit{whenever} \quad z_{-s} \in R, \quad z_s = \frac{1 - \theta_0^*}{\|\theta_1^*\|_2},$$

$$f_-(A_s^\top z) \geq c_-, \qquad \textit{whenever} \quad z_{-s} \in R, \quad z_s = \frac{-1 - \theta_0^*}{\|\theta_1^*\|_2}.$$

Assumption 6. *X_{-1} is a sub-Gaussian random vector. This implies there exists a constant $\nu > 0$ such that for any unit vector μ and any $t > 0$, $\Pr(|\mu^\top (X_{-1} - \mathbb{E}X_{-1})| > t\nu) \leq 2e^{-t^2/2}$.*

Assumption 7. *As $n \to \infty$, the sequences h_n and λ_n satisfy $h_n \to 0, \lambda_n \to 0, \dfrac{nh_n^2}{\sqrt{\log(1/h_n)}} \to \infty$ and $\sqrt{n}(h_n^2 + \lambda_n) \to 0$.*

Assumption 1 ensures randomized response tends to capture true labels (Zhang et al., 2025). Assumptions 2–6 are adapted from Rybak et al. (2025). We impose additional conditions $K(\cdot) \in C^1(\mathbb{R})$ and $\|K'(\cdot)\|_\infty < \infty$ in Assumption 2, which are satisfied by commonly used density functions. Assumptions 4 and 5 are satisfied if $X_{-1} \mid Y^* = y$ follows a Gaussian distribution for each $y \in \{-1, +1\}$. Assumption 7 ensures the bias induced by smoothing and regularization is asymptotically negligible on the $\sqrt{n}$-scale, which permits us to derive the limit theory around θ^*.

Proposition 1. *Suppose Assumptions 1 and 2 hold, and $p_{+1} < 1$ or $p_{-1} < 1$. Then for every sufficiently small $h > 0$, there exist (x, y, θ) such that $\tilde{l}_h(\theta; x, y)$ is not convex.*

Theorem 1. *Suppose Assumptions 1–7 hold. There exist $R_1 > 0$, $m_0 > 0$, and $h_0 > 0$ such that for all $0 < h \le h_0$, it holds that $\Pr\left(\inf_{\theta \in \mathcal{N}_h} \lambda_{\min}\{\nabla^2 M_n(\theta)\} \ge m_0\right) \to 1$, where $\mathcal{N}_h = B(\theta^*, hR_1)$. Furthermore, redefining*

$$\hat{\theta}_h = \arg\min_{\theta \in \mathcal{N}_h} M_n(\theta),$$

with probability tending to one, $M_n(\theta)$ possesses a unique local minimizer $\hat{\theta}_h$ in $\mathcal{N}_h$.

Although Proposition 1 implies the empirical objective $M_n(\theta)$ is not globally convex in general, Theorem 1 shows $M_n(\theta)$ is locally strongly convex in the neighborhood $\mathcal{N}_h$ with probability tending to one, which permits the proper definition of our estimator $\hat{\theta}_h$ as the unique local minimizer of $M_n(\theta)$ within $\mathcal{N}_h$.

Theorem 2. *Suppose Assumptions 1–7 hold. Let $H_0 := G(\theta^*, 0)$, where G is defined in Proposition 3. Define*

$$\zeta(\theta; X, +1) = -\frac{X}{p_{-1} + p_{+1} - 1}\{p_{-1}\mathbf{1}(1 - X^\top\theta \ge 0) + (1 - p_{+1})\mathbf{1}(1 + X^\top\theta \ge 0)\},$$

$$\zeta(\theta; X, -1) = \frac{X}{p_{-1} + p_{+1} - 1}\{(1 - p_{-1})\mathbf{1}(1 - X^\top\theta \ge 0) + p_{+1}\mathbf{1}(1 + X^\top\theta \ge 0)\},$$

and $V_0 := \mathrm{Var}(\zeta(\theta^*; X, Y))$. The local estimator $\hat{\theta}_h$ satisfies

$$\sqrt{n}(\hat{\theta}_h - \theta^*) \Rightarrow N\left(0, H_0^{-1} V_0 H_0^{-1}\right).$$

Theorem 2 shows that the asymptotic covariance matrix of $\hat{\theta}_h$ has the sandwich form $H_0^{-1} V_0 H_0^{-1}$, where H_0 does not depend on the randomized response probabilities, while V_0 depends on (p_{-1}, p_{+1}) through the corrected score function.

Theorem 3. *Suppose Assumptions 1–7 hold. Define $\hat{H}_n := \nabla^2 M_n(\hat{\theta}_h)$, $s_i(\theta) := \nabla \tilde{l}_h(\theta; X_i, Y_i)$,*

$$\hat{V}_n := \frac{1}{n}\sum_{i=1}^n \left(s_i(\hat{\theta}_h) - \frac{1}{n}\sum_{j=1}^n s_j(\hat{\theta}_h)\right)\left(s_i(\hat{\theta}_h) - \frac{1}{n}\sum_{j=1}^n s_j(\hat{\theta}_h)\right)^\top$$

and the sandwich estimator $\hat{\Sigma}_n := \hat{H}_n^{-1}\hat{V}_n\hat{H}_n^{-1}$. We have

$$\hat{H}_n \xrightarrow{p} H_0, \qquad \hat{V}_n \xrightarrow{p} V_0, \qquad \hat{\Sigma}_n \xrightarrow{p} H_0^{-1} V_0 H_0^{-1}.$$

Corollary 1. *Suppose Assumptions 1–7 hold. For any fixed nonzero vector $a \in \mathbb{R}^{d+1}$, we have $\dfrac{\sqrt{n}a^\top(\hat{\theta}_h - \theta^*)}{\sqrt{a^\top\hat{\Sigma}_n a}} \Rightarrow N(0,1)$. For the j-th coordinate, $(\hat{\theta}_h)_j \pm z_{1-\alpha/2}\sqrt{\dfrac{(\hat{\Sigma}_n)_{jj}}{n}}$ gives an asymptotic $1 - \alpha$ confidence interval.*

Theorem 3 provides a consistent estimator of the asymptotic covariance matrix, and Corollary 1 yields asymptotically valid confidence intervals for individual components of the parameter θ^*.

4 Optimal randomized response mechanism under LabelDP

In this section, we first establish an optimization framework and then derive the optimal randomized response mechanisms under ε-LabelDP and (ε, δ)-LabelDP constraints.

4.1 Optimization Objective

Differential privacy is a standard framework for protecting individual information (Dwork et al., 2006; Dwork & Roth, 2014). LabelDP specializes differential privacy to settings in which the labels are sensitive while the covariates are non-private (Chaudhuri & Hsu, 2011; Busa-Fekete et al., 2023).

Definition 1 ((ε, δ)-LabelDP (Ghazi et al., 2021))**.** *Given the privacy parameters $\varepsilon \geq 0$ and $\delta \in [0, 1)$, a randomized algorithm $\mathcal{A}$ satisfies (ε, δ)-label differential privacy if, for any two adjacent datasets D and D' that differ in the label of a single example, and for any subset of outputs $S \subseteq$ Range$(\mathcal{A})$, the following inequality holds:*

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(D') \in S] + \delta.$$

Specifically, when $\delta = 0$, the algorithm satisfies ε-LabelDP. A smaller ε indicates a stronger level of privacy protection.

Following Holohan et al. (2017), we first characterize the feasible regions of the randomized response parameters (p_{-1}, p_{+1}) under LabelDP constraints.

Definition 2. *A randomized response mechanism satisfies ε-LabelDP if $(p_{-1}, p_{+1}) \in \mathcal{R}$, or satisfies (ε, δ)-LabelDP if $(p_{-1}, p_{+1}) \in \mathcal{R}'$, where $\mathcal{R}$ and $\mathcal{R}' \subset \mathbb{R}^2$ are defined as*

$$\mathcal{R} = \left\{(p_{-1}, p_{+1}) \in [0,1]^2 : p_{-1} + p_{+1} > 1,\ p_{-1} \leq e^{\varepsilon}(1 - p_{+1}),\ p_{+1} \leq e^{\varepsilon}(1 - p_{-1})\right\},$$

$$\mathcal{R}' = \left\{(p_{-1}, p_{+1}) \in [0,1]^2 : p_{-1} + p_{+1} > 1,\ p_{-1} \leq e^{\varepsilon}(1 - p_{+1}) + \delta,\ p_{+1} \leq e^{\varepsilon}(1 - p_{-1}) + \delta\right\}.$$

To improve the statistical efficiency of estimator $\hat{\theta}_h$, we analyze the asymptotic covariance of $\hat{\theta}_h$. By Theorem 2, the asymptotic covariance matrix is $H_0^{-1} V_0 H_0^{-1}$, where H_0 is independent of the randomized response mechanism.

Proposition 2. *Suppose Assumptions 1–4 hold. Then V_0 admits the decomposition $V_0 = V_{\mathrm{np}} + V_{\mathrm{priv}}$, and*

$$V_{\mathrm{np}} = \mathrm{Var}\left(-Y^* X \mathbf{1}\{1 - Y^* X^\top \theta^* \geq 0\}\right),$$
$$V_{\mathrm{priv}} = \mathbb{E}\left[\frac{p_{+1}(1 - p_{+1})q + p_{-1}(1 - p_{-1})(1 - q)}{(p_{-1} + p_{+1} - 1)^2}(A + B)^2 X X^\top\right],$$

where $A = \mathbf{1}\{1 - X^\top \theta^* > 0\}$, $B = \mathbf{1}\{1 + X^\top \theta^* > 0\}$, $q = \Pr(Y^* = 1 \mid X)$.

The decomposition in Proposition 2 separates the asymptotic covariance into two components. V_{np} is the variance that does not depend on the randomized response mechanism. V_{priv} is the additional variance introduced by randomized response and depends on (p_{-1}, p_{+1}). Proposition 2 implies our objective simplifies to analyzing V_{priv} only, since V_{np} is independent of the randomized response mechanism.

In the LabelDP setting, the randomized response mechanism acts only on the labels. Consequently, when isolating the additional variability introduced by randomized response, it is natural to condition on the observed design $(X_1, \ldots, X_n)$. Under this conditional perspective, the covariates are treated as fixed, and the remaining randomness comes only from the conditional distribution of the true labels and from the randomized response mechanism. Thus, the conditional variance isolates the additional covariance contribution due to randomized response, without mixing it with the sampling variability of the covariates. Then conditional on the given observed design $(X_1, \ldots, X_n)$, V_{priv} admits the representation

$$V_{\mathrm{condp}} := \frac{1}{n}\sum_{i=1}^{n} \frac{p_{+1}(1 - p_{+1})q_i + p_{-1}(1 - p_{-1})(1 - q_i)}{(p_{-1} + p_{+1} - 1)^2}(A_i + B_i)^2 X_i X_i^\top,$$

where $A_i = \mathbf{1}\{1 - X_i^\top \theta^* > 0\}$, $B_i = \mathbf{1}\{1 + X_i^\top \theta^* > 0\}$, $q_i = \Pr(Y_i^* = 1 \mid X_i)$.

This conditional formulation provides a sharper and more interpretable decomposition of the variance contribution attributable specifically to privacy. V_{priv} is the population counterpart of V_{condp}: it is obtained by averaging the randomized-response variance contribution conditional on the covariates over the marginal distribution of X.

We adopt a trace criterion for improving overall inferential efficiency (Kiefer, 1974; Pukelsheim, 2006; Zhang et al., 2025). Therefore, we determine the optimal randomized response parameters by minimizing the trace of V_{condp}, denoted as $T(\theta^*; p_{-1}, p_{+1})$.

$$T\left(\theta^*; p_{-1}, p_{+1}\right) := \frac{1}{n}\sum_{i=1}^{n} \frac{p_{+1}(1-p_{+1})q_i + p_{-1}(1-p_{-1})(1-q_i)}{(p_{+1}+p_{-1}-1)^2}(A_i+B_i)^2\|X_i\|_2^2$$

We adopt the trace criterion rather than the determinant criterion for the following reasons. First, the trace of the asymptotic covariance matrix equals the sum of the asymptotic variances of the individual parameter components, which is aligned with our inferential target. Second, the trace criterion is computationally simpler. The trace criterion is numerically more stable when the covariance matrix is nearly singular, whereas the determinant can be dominated by small eigenvalues. Both criteria yield consistent estimators in asymptotic efficiency, but the trace criterion achieves this with substantially lower computational cost and greater practical tractability.

To get the optimal (p_{-1}, p_{+1}) in the feasible region, we solve the following optimization problem

$$(\hat{p}_{-1}, \hat{p}_{+1}) = \underset{(p_{-1},p_{+1})\in\mathcal{R}\text{or}\mathcal{R}'}{\arg\min} T(\theta^*; p_{-1}, p_{+1})$$

Assumption 8. *Suppose that* $0 < q_i = \Pr(Y_i^* = 1 \mid X_i) < 1$.

Assumption 8 rules out the trivial case of perfect conditional separation and ensures that each observation retains genuine label uncertainty given the covariates.

Lemma 1. *Suppose Assumptions 1 and 8 hold. There exists* (p^*_{-1}, p^*_{+1}) *such that*

$$(p^*_{-1}, p^*_{+1}) \in \underset{(p_{-1},p_{+1})\in\mathcal{R}or\mathcal{R}'}{\arg\min} T(\theta^*; p_{-1}, p_{+1}),$$

where (p^*_{-1}, p^*_{+1}) *lie on the boundaries of* $\mathcal{R}$ *or* $\mathcal{R}'$.

Lemma 1 confirms that the optimal solution (p^*_{-1}, p^*_{+1}) strictly resides on the upper boundaries of the feasible privacy regions $\mathcal{R}$ and $\mathcal{R}'$. Next we will derive the optimal LabelDP mechanisms under both ε-LabelDP and (ε, δ)-LabelDP constraints.

4.2 Optimal mechanisms under ε-LabelDP and (ε, δ)-LabelDP

Theorem 4 (ε-LabelDP)**.** *Suppose Assumptions 1 and 8 hold. For a given privacy budget* $\varepsilon > 0$*, by minimizing the objective function* $T(\theta^*; p_{-1}, p_{+1})$*, the optimal* ε*-LabelDP randomized response mechanism is given by the probabilities* $p_{-1} = p_{+1} = \dfrac{e^\varepsilon}{e^\varepsilon + 1}$.

Theorem 5 ((ε, δ)-LabelDP)**.** *Suppose Assumptions 1 and 8 hold. Given* $\varepsilon > 0$ *and* $\delta \in (0, 1)$*, by minimizing the objective function* $T(\theta^*; p_{-1}, p_{+1})$*, the optimal* (ε, δ)*-LabelDP randomized response mechanisms are described as follows:*

(i) If $T(\theta^*; \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1}, \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1}) \le T(\theta^*; 1, \delta)$ *and* $T(\theta^*; \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1}, \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1}) \le T(\theta^*; \delta, 1)$*, then the probabilities are* $p_{-1} = p_{+1} = \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1}$.

(ii) If $T(\theta^*; 1, \delta) \le T(\theta^*; \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1}, \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1})$ *and* $T(\theta^*; 1, \delta) \le T(\theta^*; \delta, 1)$*, then the probabilities are* $p_{-1} = 1$ *and* $p_{+1} = \delta$.

(iii) If $T(\theta^*; \delta, 1) \le T(\theta^*; \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1}, \dfrac{e^\varepsilon+\delta}{e^\varepsilon+1})$ *and* $T(\theta^*; \delta, 1) \le T(\theta^*; 1, \delta)$*, then the probabilities are* $p_{-1} = \delta$ *and* $p_{+1} = 1$.

Theorem 5 is not directly computable, because it depends on the unknown θ^* and q_i. To resolve this issue, we adopt the trace of the empirical covariance estimator, $\text{tr}(\hat{V}_n)$, as a data-driven surrogate criterion to evaluate and select the optimal solution that best fits the given sample. This choice is natural because Theorem 3 shows that $\hat{V}_n$ consistently estimates V_0.

In practical applications, the optimal (ε, δ)-LabelDP mechanism is determined through three steps. First, we substitute the three candidate sets of extreme probability pairs, specifically $(p_{-1}, p_{+1}) \in \{(1, \delta), (\delta, 1), ((e^{\varepsilon} + \delta)/(e^{\varepsilon} + 1), (e^{\varepsilon} + \delta)/(e^{\varepsilon} + 1))\}$, into (3) to compute the corresponding estimate $\hat{\theta}_h$ in each scenario. Next, based on these empirical estimates, we evaluate $\mathrm{tr}(\hat{V}_n)$ for each of the three parameter configurations. Finally, the candidate with the smallest $\mathrm{tr}(\hat{V}_n)$ is selected as the practical (ε, δ)-LabelDP mechanism.

It is crucial to highlight that this selection procedure does not require direct access to the underlying sensitive response data. Consequently, while strictly preserving data privacy, this approach successfully yields the optimal estimate of the SVM parameter θ^* and facilitates valid statistical inference. This mechanism ultimately empowers researchers to conduct valid statistical inference based on sensitive label data.

5 Numerical Studies

In this section, we conduct simulation studies and a real-data application to evaluate the proposed method.

5.1 Simulation

We consider a balanced Gaussian class-conditional scenario, which is both analytically transparent and fully compatible with the assumptions used in our asymptotic theory. Let the true labels be $Y_i^* \in \{-1, +1\}$ with $\mathbb{P}(Y_i^* = +1) = \mathbb{P}(Y_i^* = -1) = \frac{1}{2}$. The covariates are generated as $(X_i)_{-1} \mid Y_i^* = +1 \sim N(\mu, \Sigma)$ and $(X_i)_{-1} \mid Y_i^* = -1 \sim N(-\mu, \Sigma)$, with $\mu = (1.2,\ 0.4,\ 1.0)^\top$ and $\Sigma = \mathrm{diag}(0.8^2,\ 1.0^2,\ 0.7^2)$. By the formula of Koo et al. (2008), the corresponding population parameter is $\theta^* = (0,\ 2.165,\ 0.462,\ 2.357)^\top$.

We set the sample size to $n = 3000$ and evaluate performance over $B = 500$ independent replications. To extensively test the robustness of our method, we study the privacy budget grid $\varepsilon \in \{0.05,\ 0.06,\ 0.07,\ 0.08,\ 0.09,\ 0.10,\ 0.50,\ 0.70,\ 1.00\}$. We report results under both ε-LabelDP ($\delta = 0$) and (ε, δ)-LabelDP with $\delta = 10^{-4}$.

We consider several benchmarks for comparison with our proposed methods:

- **NP**: Nonprivate convolution-smoothed SVM, which estimates the population target without any privacy consideration.
- **RRSVM**: Applies the optimal LabelDP randomized response mechanism, then fits a convolution-smoothed SVM using the private labels (Holohan et al., 2017). The naive approach ignores the extra bias introduced by randomization.
- **ORRSVM**: Optimal LabelDP randomized response convolution-smoothed SVM (the proposed method).

For NP and RRSVM, we use a Gaussian kernel with bandwidth $h = (d/n)^{1/4}$ ($d = 3$) from Rybak et al. (2025), computing them as global minimizers. By contrast, ORRSVM is optimized locally, initialized at the RRSVM estimator, with theory-driven sequences $h_n = 0.3\, n^{-1/3}$ and $\lambda_n = 0.05\, n^{-2/3}$.

The performance is evaluated through two quantities. The first is the empirical mean squared error $\mathrm{MSE} = (\frac{1}{B}\sum_{b=1}^{B} \|\hat{\theta}^{(b)} - \theta^*\|_2^2)$, and $\hat{\theta}^{(b)}$ is the estimator obtained from the b-th replication. The second is the empirical average component-wise coverage probability of nominal 95% Wald confidence intervals, $\mathrm{ACP} = \frac{1}{B(d+1)}\sum_{b=1}^{B}\sum_{j=1}^{d+1} \mathbf{1}\{\theta_j^* \in [L_j^{(b)}, U_j^{(b)}]\}$, where $L_j^{(b)}$ and $U_j^{(b)}$ denote the lower and upper bounds, respectively, of the confidence interval for the j-th component θ_j^* constructed in the b-th Monte Carlo replication.

Figure 1 presents the empirical $\log_{10}(\mathrm{MSE})$ of $\hat{\theta}$ and the empirical average component-wise coverage probability of the approximate 95% confidence intervals under ε- and (ε, δ)-LabelDP con-

straints. Among the private procedures, the contrast between RRSVM and ORRSVM is clear. Under ε-LabelDP, the RRSVM benchmark exhibits large estimation error across the entire privacy grid, with empirical $\log_{10}(\mathrm{MSE})$ values ranging from 0.984 at $\varepsilon = 0.05$ to 0.843 at $\varepsilon = 1.00$. Its interval estimates are also badly miscalibrated, with ACP values remaining between 0.237 and 0.274.

By comparison, ORRSVM achieves uniformly lower estimation error. Its empirical $\log_{10}(\mathrm{MSE})$ decreases monotonically from 0.906 at $\varepsilon = 0.05$ to 0.406 at $\varepsilon = 1.00$, moving progressively closer to the nonprivate benchmark as the privacy budget relaxes. On the original MSE scale, ORRSVM reduces RRSVM's estimation error by approximately 17% at $\varepsilon = 0.05$, with the advantage growing to approximately 63% at $\varepsilon = 1.00$. The corrected objective restores the linear SVM parameter and the gain becomes more pronounced as the privacy noise weakens. Regarding coverage, ORRSVM's ACP ranges between 0.686 and 0.806, which is substantially higher than RRSVM's coverage.

The (ε, δ)-LabelDP results lead to the same substantive conclusions. ORRSVM continues to dominate RRSVM at every ε, with empirical $\log_{10}(\mathrm{MSE})$ values ranging from 0.914 to 0.401 and ACP values ranging from 0.713 to 0.791. The RRSVM benchmark again performs poorly, with $\log_{10}(\mathrm{MSE})$ between 0.985 and 0.843 and ACP between 0.237 and 0.271. We also note that ORRSVM's empirical coverage remains below the nominal 95% level throughout the privacy grid. Nevertheless, ORRSVM's coverage remains substantially closer to the nominal level than RRSVM's across the entire grid, confirming that correcting for the randomized response mechanism materially improves inferential validity. The key to the proposed method is that the corrected objective restores the linear SVM parameter and removes the structural bias induced by the randomized response mechanism. Consequently, the performance gain of the proposed method is most pronounced under the strongest privacy protection.

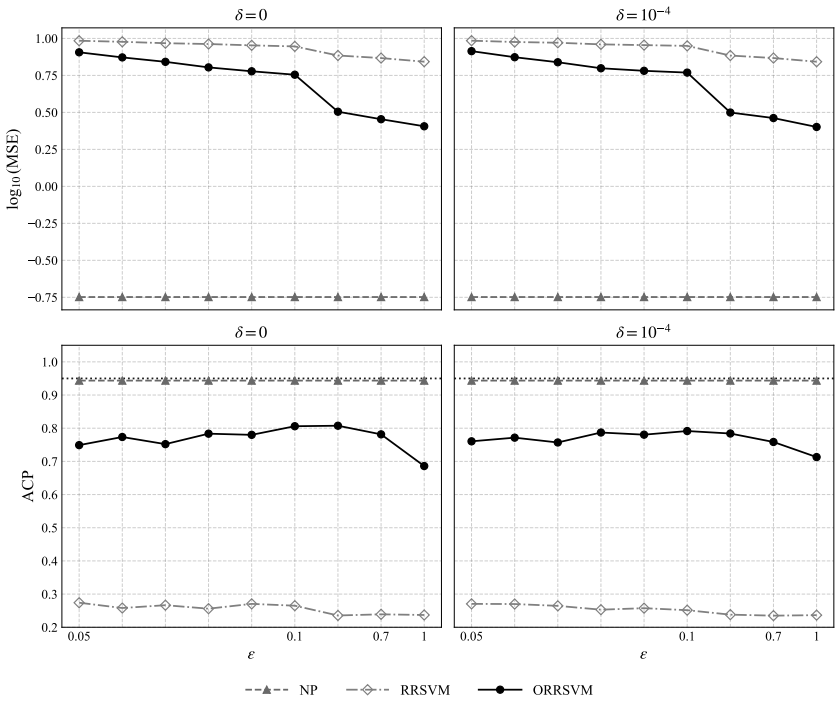

Figure 1: Empirical $\log_{10}(\mathrm{MSE})$ and ACP for three methods against different ε.

5.2 Real-data Application

We analyze the Cardiovascular Disease dataset (Ulianova, 2019) from Kaggle as a real-data experiment. The dataset consists of $70,000$ observations with clinical features, and a binary label indicating whether each patient has cardiovascular disease (+1) or not (-1). We standardize all features and then apply principal component analysis, retaining the leading 5 principal components.

The dataset is partitioned using 10 stratified random splits, allocating 70% of the data to train the model and the remaining 30% for testing. We evaluated the three methods: NP, RRSVM, and ORRSVM under the identical privacy budget grid as in our simulation study. The performance of these models was assessed by evaluating the out-of-sample classification accuracy on the test sets.

Figure 2 shows the test accuracy results under ε- and (ε, δ)-LabelDP constraints. Across the full privacy range, ORRSVM remains much closer to NP than the RRSVM benchmark. Under ε-LabelDP, ORRSVM attains average accuracy between 0.590 and 0.675, whereas RRSVM ranges from 0.528 to 0.643. Under (ε, δ)-LabelDP with $\delta = 10^{-4}$, ORRSVM ranges from 0.591 to 0.675, while RRSVM ranges from 0.538 to 0.651. The large differences appear when privacy is strong. For example, at $\varepsilon = 0.05$, ORRSVM achieves about 0.590 accuracy under ε-LabelDP and 0.624 under (ε, δ)-LabelDP, compared with 0.538 and 0.538 for RRSVM. As ε increases, both private procedures improve, but ORRSVM maintains a clear advantage throughout.

The real-data application supports the same conclusion as the simulation study. When directly fitting the SVM to privatized labels without correcting for the randomized response mechanism, prediction accuracy deteriorates substantially at small ε. Once the correction is accounted for, much of the NP predictive performance is retained and the performance gap between ORRSVM and the non-private benchmark decreases as ε increases.

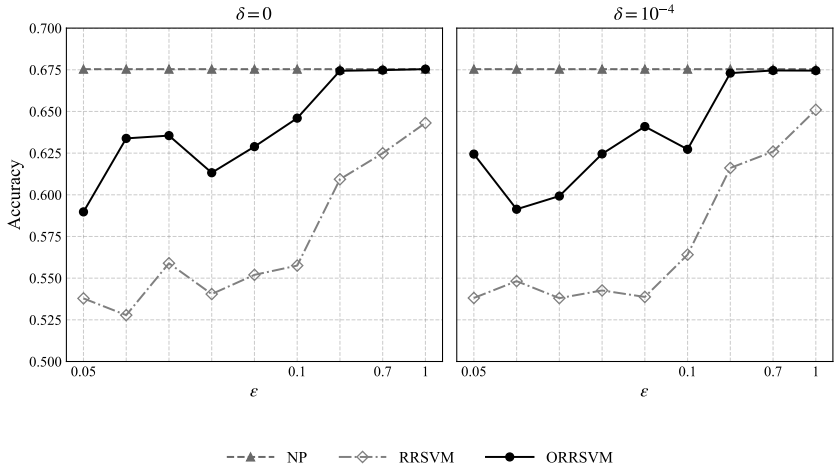

Figure 2: Accuracy for three methods against different ε.

6 Conclusion

This paper studies statistical inference for the linear SVM under LabelDP. We first construct a corrected loss function and apply convolution smoothing to the hinge loss to obtain a differentiable empirical objective. We then establish the asymptotic normality of the estimator and construct a consistent covariance estimator, yielding asymptotically valid confidence intervals. We further derive optimal randomized response mechanisms under both ε-LabelDP and (ε, δ)-LabelDP constraints, thereby directly linking mechanism design to statistical efficiency. Numerical studies reveal that the proposed method outperforms a naive randomized response SVM benchmark in parameter estimation, confidence interval coverage, and prediction accuracy.

The present work also opens several directions for future research. A first direction is to extend the proposed framework to multiclass SVM under k-ary randomized response (Kairouz et al., 2016; Ye & Barg, 2017). A second is to move beyond the current fixed-dimensional setting and study high-dimensional sparse SVMs. Finally, it would be valuable to develop analogous theory under Rényi or Gaussian privacy constraints (Mironov, 2017; Dong et al., 2022).

AI USE STATEMENT

In this work, we used generative AI tools solely to polish the writing of the manuscript and we have carefully reviewed all AI-assisted work. We have not used generative AI tools for research ideas, theoretical claims, or experiments. We take responsibility for the final content of this work, including text with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we have submitted the anonymized source code implementing simulation and real-data experiments as downloadable supplementary material.

REFERENCES

Liaqat Ali, Awais Niamat, Javed Ali Khan, Noorbakhsh Amiri Golilarz, Xiong Xingzhong, Adeeb Noor, Redhwan Nour, and Syed Ahmad Chan Bukhari. An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access*, 7:54007–54014, 2019. doi: 10.1109/ACCESS.2019.2909969.

Graeme Blair, Kosuke Imai, and Yang-Yang Zhou. Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, 110(511):1304–1319, July 2015. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2015.1050028.

Robert Istvan Busa-Fekete, Andres Munoz Medina, Umar Syed, and Sergei Vassilvitskii. Label differential privacy and private training data release. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 3233–3251. PMLR, July 2023.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 155–186. JMLR Workshop and Conference Proceedings, December 2011.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. ISSN 1573-0565. doi: 10.1007/BF00994018.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 84(1):3–37, February 2022. ISSN 1369-7412. doi: 10.1111/rssb.12454.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–487, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-32732-5. doi: 10.1007/11681878_14.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27131–27145. Curran Associates, Inc., 2021.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, January 2002. ISSN 1573-0565. doi: 10.1023/A:1012487302797.

Naoise Holohan, Douglas J. Leith, and Oliver Mason. Optimal differentially private mechanisms for randomised response. *IEEE Transactions on Information Forensics and Security*, 12(11): 2726–2735, November 2017. ISSN 1556-6013, 1556-6021. doi: 10.1109/TIFS.2017.2718487.

Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4):847–856, 2007. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2006.07.007. URL https://www.sciencedirect.com/science/article/pii/S095741740600217X.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Journal of Machine Learning Research*, 17(17):1–51, 2016.

Jack Kiefer. General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2(5):849–879, 1974. doi: 10.1214/aos/1176342810.

Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9(44):1343–1368, 2008.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, August 2017. doi: 10.1109/CSF.2017.11.

Friedrich Pukelsheim. *Optimal Design of Experiments*. SIAM, 2006. doi: 10.1137/1.9780898719109.

Jakub Rybak, Heather Battey, and Wen-Xin Zhou. On inference for the support vector machine. *Journal of Machine Learning Research*, 26(85):1–54, 2025.

Xinyu Tang, Milad Nasr, Saeed Mahloujifar, Virat Shejwalkar, Liwei Song, Amir Houmansadr, and Prateek Mittal. Machine learning with differentially private labels: mechanisms and frameworks. *Proceedings on Privacy Enhancing Technologies*, 2022. ISSN 2299-0984. doi: 10.56553/popets-2022-0112.

Svetlana Ulianova. Cardiovascular disease dataset, 2019. URL `https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset`.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, March 1965. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1965.10480775.

Shirong Xu, Chendi Wang, Will Wei Sun, and Guang Cheng. Binary classification under local label differential privacy using randomized response mechanisms. *Transactions on Machine Learning Research*, 2023, 2023.

Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under local differential privacy. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 759–763, June 2017. doi: 10.1109/ISIT.2017.8006630.

Ling-Li Zeng, Hui Shen, Li Liu, Lubin Wang, Baojuan Li, Peng Fang, Zongtan Zhou, Yaming Li, and Dewen Hu. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain*, 135(5):1498–1507, 05 2012. ISSN 0006-8950. doi: 10.1093/brain/aws059.

Ce Zhang, Yixin Han, Yafei Wang, Xiaodong Yan, Linglong Kong, Ting Li, and Bei Jiang. Differentially private analysis for binary response models: Optimality, estimation, and inference. In *Forty-second International Conference on Machine Learning*, 2025.

A APPENDIX

Organization. Appendix A.1 introduces analytic preliminaries. Appendix A.2 proves that the corrected smoothed loss is generally not globally convex. Appendix A.3 establishes the local strong convexity of the empirical objective and proves the existence and uniqueness of the localized estimator. Appendix A.4 and Appendix A.5 prove the asymptotic normality of the estimator and the consistency of the feasible covariance estimator, leading to confidence intervals based only on privatized labels. Appendix A.6 provides the variance decomposition induced by randomized response and prove the optimal LabelDP mechanisms under both ε-LabelDP and (ε, δ)-LabelDP constraints.

A.1 ANALYTIC PRELIMINARIES

Proposition 3. *Let $\Delta := p_{+1} + p_{-1} - 1 > 0$, $\bar{K}(t) := \int_{-\infty}^{t} K(v)\,dv$, $\mathcal{N}_n := \mathcal{N}_{h_n} = B(\theta^*, h_n R_1)$. Under Assumptions 1–4, the following statements hold.*

(i) For every (x, y), the map $\theta \mapsto \tilde{l}_h(\theta; x, y)$ is twice continuously differentiable, and

$$\nabla \tilde{l}_h(\theta; x, +1) = -\frac{x}{\Delta}\left[p_{-1}\bar{K}\left(\frac{1 - x^\top\theta}{h}\right) + (1 - p_{+1})\bar{K}\left(\frac{1 + x^\top\theta}{h}\right)\right], \tag{4}$$

$$\nabla^2 \tilde{l}_h(\theta; x, +1) = \frac{xx^\top}{h\Delta}\left[p_{-1}K\left(\frac{1 - x^\top\theta}{h}\right) - (1 - p_{+1})K\left(\frac{1 + x^\top\theta}{h}\right)\right], \tag{5}$$

$$\nabla \tilde{l}_h(\theta; x, -1) = \frac{x}{\Delta}\left[(1 - p_{-1})\bar{K}\left(\frac{1 - x^\top\theta}{h}\right) + p_{+1}\bar{K}\left(\frac{1 + x^\top\theta}{h}\right)\right], \tag{6}$$

$$\nabla^2 \tilde{l}_h(\theta; x, -1) = \frac{xx^\top}{h\Delta}\left[p_{+1}K\left(\frac{1 + x^\top\theta}{h}\right) - (1 - p_{-1})K\left(\frac{1 - x^\top\theta}{h}\right)\right]. \tag{7}$$

(ii) Let $M(\theta) := \mathbb{E}\tilde{l}_h(\theta; X, Y) + \frac{\lambda}{2}\theta^\top J\theta$,$H(\theta) := \nabla^2 M(\theta)$, $L_h^*(\theta) := \mathbb{E}\, l_h(1 - Y^* X^\top\theta)$. For every θ, almost surely,

$$\mathbb{E}\big[\tilde{l}_h(\theta; X, Y) \mid X, Y^*\big] = l_h\big(1 - Y^* X^\top\theta\big), \tag{8}$$

$$\mathbb{E}\big[\nabla \tilde{l}_h(\theta; X, Y) \mid X, Y^*\big] = -Y^* X\, \bar{K}\left(\frac{1 - Y^* X^\top\theta}{h}\right), \tag{9}$$

$$\mathbb{E}\big[\nabla^2 \tilde{l}_h(\theta; X, Y) \mid X, Y^*\big] = \frac{XX^\top}{h} K\left(\frac{1 - Y^* X^\top\theta}{h}\right). \tag{10}$$

Consequently,

$$M(\theta) = L_h^*(\theta) + \frac{\lambda}{2}\theta^\top J\theta. \tag{11}$$

(iii) Define, for $\theta \in \mathcal{N}$, $u \in \mathbb{R}$, and $y \in \{-1, +1\}$,

$$\chi_y((x_{-1})_{-s}, u; \theta) := \frac{1 - y\theta_0 - y(\theta_1)_{-s}^\top (x_{-1})_{-s} - u}{y\theta_{1s}}.$$

Then the functions

$$G(\theta, u) := \sum_{y=\pm 1} \frac{\pi_y}{|\theta_{1s}|} \int_{\mathbb{R}^{d-1}} x\big((x_{-1})_{-s}, \chi_y((x_{-1})_{-s}, u; \theta)\big) x\big((x_{-1})_{-s}, \chi_y((x_{-1})_{-s}, u; \theta)\big)^\top \tag{12}$$

$$\times\, f_y\big((x_{-1})_{-s}, \chi_y((x_{-1})_{-s}, u; \theta)\big)\, d(x_{-1})_{-s}, \tag{13}$$

$$g(\theta, u) := \sum_{y=\pm 1} \frac{y\pi_y}{|\theta_{1s}|} \int_{\mathbb{R}^{d-1}} x\big((x_{-1})_{-s}, \chi_y((x_{-1})_{-s}, u; \theta)\big) \tag{14}$$

$$\times\, f_y\big((x_{-1})_{-s}, \chi_y((x_{-1})_{-s}, u; \theta)\big)\, d(x_{-1})_{-s}. \tag{15}$$

are well defined on $\mathcal{N} \times \mathbb{R}$ and satisfy, for every bounded measurable φ,

$$\mathbb{E}\big[XX^\top \varphi(1 - Y^* X^\top \theta)\big] = \int_{\mathbb{R}} G(\theta, u)\varphi(u)\, du, \tag{16}$$

$$\mathbb{E}\big[Y^* X \varphi(1 - Y^* X^\top \theta)\big] = \int_{\mathbb{R}} g(\theta, u)\varphi(u)\, du. \tag{17}$$

Moreover,

$$\sup_{(\theta,u)\in\mathcal{N}\times\mathbb{R}} \|G(\theta, u)\|_{\mathrm{op}} < \infty, \qquad \sup_{(\theta,u)\in\mathcal{N}\times\mathbb{R}} \|\partial_u g(\theta, u)\|_2 < \infty. \tag{18}$$

(iv) For every $\theta \in \mathcal{N}$, every $\sigma \in \{\pm 1\}$, and every $y \in \{\pm 1\}$,

$$\mathbb{P}\big(1 - \sigma X^\top \theta = 0 \mid Y^* = y\big) = 0. \tag{19}$$

(v) Let $L(\theta) := \mathbb{E}(1 - Y^* X^\top \theta)_+$. Consequently, for every $\theta \in \mathcal{N}$,

$$\nabla L(\theta) = -\mathbb{E}\big[Y^* X \mathbf{1}\{1 - Y^* X^\top \theta \geq 0\}\big] = -\int_{\mathbb{R}} g(\theta, u)\mathbf{1}\{u \geq 0\}\, du, \tag{20}$$

$$\nabla M(\theta) = -\int_{\mathbb{R}} g(\theta, u)\bar{K}(u/h)\, du + \lambda J\theta, \tag{21}$$

$$H(\theta) = \int_{\mathbb{R}} G(\theta, u)\frac{1}{h}K(u/h)\, du + \lambda J = \int_{\mathbb{R}} G(\theta, hv)K(v)\, dv + \lambda J. \tag{22}$$

Proof. Part (i) follows standard differentiation rules.

For part (ii), consider first $Y^* = +1$. Using the conditional law $\mathbb{P}(Y = +1 \mid Y^* = +1) = p_{+1}$,

$$\begin{aligned}\mathbb{E}\big[\tilde{l}_h(\theta; X, Y) \mid X, Y^* = +1\big] &= \frac{p_{+1}\{p_{-1} l_h(1 - X^\top\theta) - (1 - p_{+1}) l_h(1 + X^\top\theta)\}}{\Delta} \\ &\quad + \frac{(1 - p_{+1})\{p_{+1} l_h(1 + X^\top\theta) - (1 - p_{-1}) l_h(1 - X^\top\theta)\}}{\Delta} \\ &= l_h(1 - X^\top\theta).\end{aligned}$$

The case $Y^* = -1$ is identical and gives (8). Taking derivatives in the two identities above, or equivalently averaging (4)–(7) with the RR probabilities, yields (9)–(10). (11) follows by taking expectations.

For part (iii), fix $y \in \{\pm 1\}$ and let φ be bounded and measurable. Conditioning on $Y^* = y$ gives

$$\mathbb{E}\big[XX^\top \varphi(1 - yX^\top\theta) \mid Y^* = y\big] = \int_{\mathbb{R}^d} x(x_{-1})x(x_{-1})^\top \varphi(1 - yX^\top\theta) f_y(x_{-1})\, dx_{-1}.$$

Write $x_{-1} = ((x_{-1})_{-s}, (x_{-1})_s)$ and use the change of variables

$$u = 1 - y\theta_0 - y(\theta_1)_{-s}^\top (x_{-1})_{-s} - y\theta_{1s}(x_{-1})_s,$$

$$(x_{-1})_s = \chi_y((x_{-1})_{-s}, u; \theta), \qquad \left|\frac{\partial u}{\partial (x_{-1})_s}\right| = |\theta_{1s}|.$$

Because of $\theta_{1s}^* \neq 0$, the Jacobian is bounded away from zero on $\mathcal{N}$, and the change of variables yields (16); the identity (17) is identical, with $y\, x(\cdot)$ in place of $x(\cdot)x(\cdot)^\top$. The envelope condition in Assumption 4 and $\theta_{1s}^* \neq 0$ then imply the uniform bounds in (18) by dominated convergence and differentiation under the integral sign.

For part (iv), conditional on $Y^* = y$, the vector X_{-1} has Lebesgue density f_y. Since $|\theta_{1s}| \geq c_s > 0$ on $\mathcal{N}$, the event $\{1 - \sigma X^\top\theta = 0\}$ is the graph of the measurable function

$$(X_{-1})_s = \chi_\sigma((X_{-1})_{-s}, 0; \theta),$$

which has Lebesgue measure zero in $\mathbb{R}^d$; hence (19) follows.

Finally, part (v) follows from parts (ii)–(iv). For (20), differentiate $L(\theta) = \mathbb{E}(1 - Y^* X^\top\theta)_+$ under the expectation. Because

$$\partial_\theta (1 - Y^* X^\top\theta)_+ = -Y^* X\, \mathbf{1}\{1 - Y^* X^\top\theta > 0\},$$

and $\mathbb{E}\|X\|_2 < \infty$, dominated convergence applies; (19) then removes the distinction between $>$ and $\geq$. (21)–(22) are immediate from (9)–(10) and (16)–(17). □

A.2 Proof of Proposition 1

Proof. Assume first that $1 - p_{+1} > 0$. By (5),

$$\nabla^2 \tilde{l}_h(\theta; x, +1) = \frac{xx^\top}{h\Delta}\left[p_{-1}K\left(\frac{1 - x^\top\theta}{h}\right) - (1 - p_{+1})K\left(\frac{1 + x^\top\theta}{h}\right)\right].$$

Choose any nonzero x and then choose θ so that $x^\top\theta = -1$. Then

$$\nabla^2 \tilde{l}_h(\theta; x, +1) = \frac{xx^\top}{h\Delta}\big[p_{-1}K(2/h) - (1 - p_{+1})K(0)\big].$$

Because $K(0) > 0$ and $K(2/h) \to 0$ as $h \downarrow 0$, the bracketed term is negative for all sufficiently small h, so the second directional derivative along x is negative. Hence $\theta \mapsto \tilde{l}_h(\theta; x, +1)$ is not convex.

If instead $1 - p_{+1} = 0$, then Assumption 1 forces $1 - p_{-1} > 0$. In that case, use (7) and choose θ so that $x^\top\theta = 1$:

$$\nabla^2 \tilde{l}_h(\theta; x, -1) = \frac{xx^\top}{h\Delta}\big[p_{+1}K(2/h) - (1 - p_{-1})K(0)\big],$$

which is again negative for all sufficiently small h. Therefore the corrected sample term is generally nonconvex whenever the RR mechanism is nontrivial. □

A.3 Proof of Theorem 1

Proof. Let $H_n(\theta) := \nabla^2 M_n(\theta)$. We work throughout with the sequence (h_n, λ_n) from Assumption 7 and abbreviate $\mathcal{N}_n := \mathcal{N}_{h_n} = B(\theta^*, h_n R_1)$.

Step 1: population geometry and the localized population center. By Proposition 14 of Rybak et al. (2025), there exist constants $R_1 > 0$, $c_0 > 0$, and $h_0 > 0$ such that for every $0 < h \le h_0$,

$$\lambda_{\min}\big(\nabla^2 L_h^*(\theta)\big) \ge c_0, \qquad \theta \in \mathcal{N}_h. \tag{23}$$

Since $H(\theta) = \nabla^2 L_h^*(\theta) + \lambda J$ by (22),

$$\lambda_{\min}\big(H(\theta)\big) \ge c_0, \qquad \theta \in \mathcal{N}_h. \tag{24}$$

In particular, M is strongly convex on $\mathcal{N}_h$.

Next we show that the smoothed population score at θ^* is of order h^2. Since $\nabla L(\theta^*) = 0$ by Assumption 3 and (20)–(21) gives

$$\nabla L_h^*(\theta^*) = -\int_{\mathbb{R}} g(\theta^*, u)\big[\bar{K}(u/h) - \mathbf{1}\{u \ge 0\}\big]du.$$

Let

$$D(v) := \bar{K}(v) - \mathbf{1}\{v \ge 0\}.$$

After the change of variables $u = hv$,

$$\nabla L_h^*(\theta^*) = -h\int_{\mathbb{R}} g(\theta^*, hv)D(v)\,dv. \tag{25}$$

Because K is symmetric, D is odd and therefore $\int_{\mathbb{R}} D(v)\,dv = 0$. Hence

$$\nabla L_h^*(\theta^*) = -h\int_{\mathbb{R}} \big[g(\theta^*, hv) - g(\theta^*, 0)\big]D(v)\,dv. \tag{26}$$

By (18),

$$\|g(\theta^*, hv) - g(\theta^*, 0)\|_2 \le Ch|v|,$$

so

$$\|\nabla L_h^*(\theta^*)\|_2 \le Ch^2\int_{\mathbb{R}} |v|\,|D(v)|\,dv = O(h^2). \tag{27}$$

The integral is finite by Assumption 2.

For each n, let

$$\tilde{\theta} := \arg\min_{\theta \in \mathcal{N}_n} M(\theta). \tag{28}$$

Existence follows from continuity, and uniqueness follows from the strong convexity in (24). Moreover,

$$\nabla M(\theta^*) = \nabla L^*_{h_n}(\theta^*) + \lambda_n J\theta^*, \qquad \|\nabla M(\theta^*)\|_2 = O(h_n^2 + \lambda_n) \tag{29}$$

by (27). If $\omega \in \mathbb{R}^{d+1}$ satisfies $\|\omega\|_2 = h_n R_1$, then $\theta^* + \omega \in \partial\mathcal{N}_n$ and strong convexity yields

$$M(\theta^* + \omega) \geq M(\theta^*) + \nabla M(\theta^*)^\top \omega + \frac{c_0}{2}\|\omega\|_2^2.$$

Therefore,

$$M(\theta^* + \omega) - M(\theta^*) \geq -\|\nabla M(\theta^*)\|_2 h_n R_1 + \frac{c_0}{2} h_n^2 R_1^2.$$

Assumption 7 implies

$$\frac{h_n^2 + \lambda_n}{h_n} = \frac{\sqrt{n}\,(h_n^2 + \lambda_n)}{\sqrt{n}\,h_n} \to 0, \tag{30}$$

because $\sqrt{n}\,(h_n^2 + \lambda_n) \to 0$ and $\sqrt{n}\,h_n \to \infty$ follows from $nh_n^2/\log(1/h_n) \to \infty$. Hence the right-hand side above is strictly positive for all large n, so the unique minimizer lies in the interior of $\mathcal{N}_n$ and satisfies

$$\nabla M(\tilde{\theta}) = 0. \tag{31}$$

Finally, using strong monotonicity and (29),

$$c_0\|\tilde{\theta} - \theta^*\|_2^2 \leq -\nabla M(\theta^*)^\top(\tilde{\theta} - \theta^*),$$

which gives

$$\|\tilde{\theta} - \theta^*\|_2 = O(h_n^2 + \lambda_n) = o(h_n). \tag{32}$$

In particular, $\tilde{\theta} \to \theta^*$.

Step 2: uniform control of the empirical Hessian. For fixed $\theta \in \mathcal{N}$ and fixed coordinates $j, k \in \{0, \ldots, d\}$, (5) and (7) imply

$$\left|e_j^\top \nabla^2 \tilde{l}_{h_n}(\theta; X, Y) e_k\right| \leq \frac{C}{h_n}|X_j X_k|. \tag{33}$$

Because X_{-1} is sub-Gaussian by Assumption 6, the augmented vector X is also sub-Gaussian, and therefore $X_j X_k$ is sub-exponential. Hence the centered quantity

$$W_{i,jk}(\theta) := e_j^\top \Big(\nabla^2 \tilde{l}_{h_n}(\theta; X_i, Y_i) - \mathbb{E}\nabla^2 \tilde{l}_{h_n}(\theta; X, Y)\Big) e_k$$

satisfies

$$\|W_{i,jk}(\theta)\|_{\psi_1} \leq C/h_n. \tag{34}$$

Bernstein's inequality therefore gives, for every fixed θ and every $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n W_{i,jk}(\theta)\right| > t\right) \leq 2\exp\big[-cn\min(h_n^2 t^2, h_n t)\big]. \tag{35}$$

Next, by the mean-value theorem, the boundedness of K', and (5), (7),

$$\big\|\nabla^2 \tilde{l}_h(\theta; x, y) - \nabla^2 \tilde{l}_h(\theta'; x, y)\big\|_{\mathrm{op}} \leq \frac{C}{h^2}\|x\|_2^3\,\|\theta - \theta'\|_2. \tag{36}$$

Consequently,

$$\|H_n(\theta) - H_n(\theta')\|_{\mathrm{op}} \leq \frac{C}{h_n^2}\left(\frac{1}{n}\sum_{i=1}^n \|x_i\|_2^3\right)\|\theta - \theta'\|_2, \tag{37}$$

$$\|H(\theta) - H(\theta')\|_{\mathrm{op}} \leq \frac{C}{h^2}\,\mathbb{E}\|X\|_2^3\,\|\theta - \theta'\|_2. \tag{38}$$

Now cover $\mathcal{N}_n = B(\theta^*, h_n R_1)$ by an η_n-net $\mathcal{G}_n$ with $\eta_n := h_n^3$.

$$\log |\mathcal{G}_n| = O\big(\log(1/h_n)\big). \tag{39}$$

Take

$$t_n := A\sqrt{\frac{\log(1/h_n)}{nh_n^2}}, \tag{40}$$

with A large enough. Since $h_n t_n \to 0$, (35) is in its quadratic regime, and a union bound over the fixed matrix coordinates and the net points gives

$$\max_{\theta_g \in \mathcal{G}_n} \|H_n(\theta_g) - H(\theta_g)\|_{\mathrm{op}} = O_p(t_n). \tag{41}$$

For any $\theta \in \mathcal{N}_n$, choose $\theta_g \in \mathcal{G}_n$ with $\|\theta - \theta_g\|_2 \le \eta_n$. Then

$$\|H_n(\theta) - H(\theta)\|_{\mathrm{op}} \le \|H_n(\theta_g) - H(\theta_g)\|_{\mathrm{op}} + \|H_n(\theta) - H_n(\theta_g)\|_{\mathrm{op}} + \|H(\theta) - H(\theta_g)\|_{\mathrm{op}}.$$

The first term is $O_p(t_n)$ by (41); the second is $O_p(h_n)$ by (37) and the law of large numbers; and the third is $O(h_n)$ by (38). Hence

$$\sup_{\theta \in \mathcal{N}_n} \|H_n(\theta) - H(\theta)\|_{\mathrm{op}} = O_p\left(\sqrt{\frac{\log(1/h_n)}{nh_n^2}} + h_n\right) = o_p(1). \tag{42}$$

Combining (24) and (42) with Weyl's inequality yields

$$\inf_{\theta \in \mathcal{N}_n} \lambda_{\min}\big(H_n(\theta)\big) \ge m_0 := c_0/2 \tag{43}$$

on an event E_n with $\mathbb{P}(E_n) \to 1$.

Step 3: existence and uniqueness of the localized empirical minimizer. Continuity of M_n on the compact set $\mathcal{N}_n$ implies that

$$\hat{\theta}_h := \arg\min_{\theta \in \mathcal{N}_n} M_n(\theta) \tag{44}$$

exists. On E_n, (43) shows that M_n is strongly convex on $\mathcal{N}_n$, so the minimizer is unique.

To prove that $\hat{\theta}_h$ lies in the interior of $\mathcal{N}_n$, define

$$S_n := \nabla M_n(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \psi_{n,i}, \qquad \psi_{n,i} := \nabla \tilde{l}_{h_n}(\tilde{\theta}; x_i, y_i) - \mathbb{E}\nabla \tilde{l}_{h_n}(\tilde{\theta}; X, Y). \tag{45}$$

By (4), (6), and $|\bar{K}| \le 1$,

$$\|\nabla \tilde{l}_{h_n}(\theta; x, y)\|_2 \le C\|x\|_2. \tag{46}$$

Hence $\mathbb{E}\|S_n\|_2^2 = O(n^{-1})$, so

$$\|S_n\|_2 = O_p(n^{-1/2}). \tag{47}$$

On E_n, strong convexity gives, for every Δ with $\tilde{\theta} + \Delta \in \mathcal{N}_n$,

$$M_n(\tilde{\theta} + \Delta) \ge M_n(\tilde{\theta}) + S_n^\top \Delta + \frac{m_0}{2}\|\Delta\|_2^2. \tag{48}$$

Choose

$$r_n := \frac{4\|S_n\|_2}{m_0} = O_p(n^{-1/2}). \tag{49}$$

Then for every $\|\Delta\|_2 = r_n$, (48) yields

$$M_n(\tilde{\theta} + \Delta) - M_n(\tilde{\theta}) \ge -\|S_n\|_2 r_n + \frac{m_0}{2} r_n^2 = \frac{m_0}{4} r_n^2 > 0. \tag{50}$$

By (32), $\mathrm{dist}(\tilde{\theta}, \partial\mathcal{N}_n) = h_n R_1 + o(h_n)$, while (49) and $\sqrt{n}\, h_n \to \infty$ imply $r_n = o_p(h_n)$. Hence $B(\tilde{\theta}, r_n) \subset \mathcal{N}_n$ with probability tending to one. On that event, (50) shows that the unique minimizer $\hat{\theta}_h$ lies inside $B(\tilde{\theta}, r_n)$, and therefore

$$\hat{\theta}_h \in \mathrm{int}(\mathcal{N}_n), \qquad \nabla M_n(\hat{\theta}_h) = 0. \tag{51}$$

□

A.4 Proof of Theorem 2

Proof. We continue to write

$$\psi_{n,i} := \nabla \tilde{l}_{h_n}(\tilde{\theta}; x_i, y_i) - \mathbb{E}\nabla \tilde{l}_{h_n}(\tilde{\theta}; X, Y), \qquad \Delta_n := \hat{\theta}_h - \tilde{\theta}.$$

From Step 3 in the proof of Theorem 1,

$$\|\Delta_n\|_2 = O_p(n^{-1/2}). \tag{52}$$

Step 1: Bahadur expansion around $\tilde{\theta}$. A Taylor expansion around $\tilde{\theta}$ gives

$$0 = \nabla M_n(\hat{\theta}_h) = \nabla M_n(\tilde{\theta}) + \bar{H}_n \Delta_n, \qquad \bar{H}_n := \int_0^1 H_n(\tilde{\theta} + t\Delta_n)\, dt. \tag{53}$$

By (45),

$$\nabla M_n(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^n \psi_{n,i}. \tag{54}$$

Hence

$$\Delta_n = -\bar{H}_n^{-1}\frac{1}{n}\sum_{i=1}^n \psi_{n,i}. \tag{55}$$

It remains to show that $\bar{H}_n = H(\tilde{\theta}) + o_p(1)$. Decompose

$$\bar{H}_n - H(\tilde{\theta}) = \int_0^1 \big[H_n(\tilde{\theta} + t\Delta_n) - H(\tilde{\theta} + t\Delta_n)\big]dt + \int_0^1 \big[H(\tilde{\theta} + t\Delta_n) - H(\tilde{\theta})\big]dt. \tag{56}$$

The first term is $o_p(1)$ by (42), since $\tilde{\theta} + t\Delta_n \in \mathcal{N}_n$ with probability tending to one. For the second term, use (22): along any subsequence we may extract a further subsequence on which $\tilde{\theta} \to \theta^*$ and $\Delta_n \to 0$ almost surely. For each fixed $t \in [0,1]$ and $v \in \mathbb{R}$,

$$G(\tilde{\theta} + t\Delta_n, h_n v) - G(\tilde{\theta}, h_n v) \to 0 \qquad \text{a.s.}$$

by continuity of G, while (18) yields the integrable domination

$$\|G(\tilde{\theta} + t\Delta_n, h_n v) - G(\tilde{\theta}, h_n v)\|_{\mathrm{op}} K(v) \le 2 \sup_{(\theta,u)\in\mathcal{N}\times\mathbb{R}} \|G(\theta, u)\|_{\mathrm{op}} K(v).$$

Dominated convergence, first in v and then in t, therefore gives

$$\int_0^1 \big[H(\tilde{\theta} + t\Delta_n) - H(\tilde{\theta})\big]dt = o_p(1). \tag{57}$$

Combining (56) and (57),

$$\bar{H}_n = H(\tilde{\theta}) + o_p(1). \tag{58}$$

Next, (22) implies

$$H(\tilde{\theta}) = \int_{\mathbb{R}} G(\tilde{\theta}, h_n v) K(v)\, dv + \lambda_n J. \tag{59}$$

Since $\tilde{\theta} \to \theta^*$, $h_n \to 0$, and $\lambda_n \to 0$, dominated convergence and (18) yield

$$H(\tilde{\theta}) \to H_0 := G(\theta^*, 0). \tag{60}$$

Moreover, taking $\theta = \theta^*$ in (23) and passing to the limit in the same way shows that

$$H_0 \succ 0. \tag{61}$$

Hence $\bar{H}_n$ is invertible with probability tending to one and

$$\bar{H}_n^{-1} = H(\tilde{\theta})^{-1} + o_p(1). \tag{62}$$

Substituting (62) into (55) gives the Bahadur expansion

$$\sqrt{n}\,(\hat{\theta}_h - \tilde{\theta}) = -H(\tilde{\theta})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{n,i} + o_p(1). \tag{63}$$

Step 2: limit score and triangular-array central limit theorem. Define

$$\zeta(\theta; X, +1) := -\frac{X}{\Delta}\Big[p_{-1}\mathbf{1}\{1 - X^\top\theta \geq 0\} + (1 - p_{+1})\mathbf{1}\{1 + X^\top\theta \geq 0\}\Big], \tag{64}$$

$$\zeta(\theta; X, -1) := \frac{X}{\Delta}\Big[(1 - p_{-1})\mathbf{1}\{1 - X^\top\theta \geq 0\} + p_{+1}\mathbf{1}\{1 + X^\top\theta \geq 0\}\Big], \tag{65}$$

and write $\zeta(\theta; X, Y)$ for the label-indexed version. Let

$$V_0 := \mathrm{Var}\big(\zeta(\theta^*; X, Y)\big). \tag{66}$$

By (46),

$$\|\nabla\tilde{l}_{h_n}(\tilde{\theta}; X, Y)\|_2 \leq C\|X\|_2. \tag{67}$$

Since $\tilde{\theta} \to \theta^*$, $h_n \to 0$, and the separating-hyperplane events have probability zero by (19), the pointwise limit of the smoothed score is

$$\nabla\tilde{l}_{h_n}(\tilde{\theta}; X, Y) \to \zeta(\theta^*; X, Y) \qquad \text{a.s.} \tag{68}$$

Dominated convergence using (67) shows that

$$\mathrm{Var}\big(\nabla\tilde{l}_{h_n}(\tilde{\theta}; X, Y)\big) \to V_0. \tag{69}$$

Now

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{n,i}\right) = \mathrm{Var}(\psi_{n,1}) \to V_0.$$

For Lindeberg's condition, (67) yields

$$\|\psi_{n,i}\|_2 \leq \|\nabla\tilde{l}_{h_n}(\tilde{\theta}; x_i, y_i)\|_2 + \mathbb{E}\|\nabla\tilde{l}_{h_n}(\tilde{\theta}; X, Y)\|_2 \leq C\|x_i\|_2 + C\mathbb{E}\|X\|_2.$$

Since $\mathbb{E}\|X\|_2^2 < \infty$, dominated convergence implies that, for every $\varepsilon > 0$,

$$\sum_{i=1}^{n}\mathbb{E}\Big[\|n^{-1/2}\psi_{n,i}\|_2^2\mathbf{1}\{\|n^{-1/2}\psi_{n,i}\|_2 > \varepsilon\}\Big] \to 0.$$

Therefore,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{n,i} \Rightarrow N(0, V_0). \tag{70}$$

Combining (63), (60), (61), and (70) gives

$$\sqrt{n}\,(\hat{\theta}_h - \tilde{\theta}) \Rightarrow N\big(0, H_0^{-1}V_0H_0^{-1}\big). \tag{71}$$

Step 3: recentering from $\tilde{\theta}$ to θ^*. By (32) and Assumption 7,

$$\sqrt{n}\,(\tilde{\theta} - \theta^*) = O\big(\sqrt{n}\,(h_n^2 + \lambda_n)\big) = o(1). \tag{72}$$

Therefore,

$$\sqrt{n}\,(\hat{\theta}_h - \theta^*) = \sqrt{n}\,(\hat{\theta}_h - \tilde{\theta}) + \sqrt{n}\,(\tilde{\theta} - \theta^*),$$

and Theorem 2 follows from (71), (72), and Slutsky theorem. □

A.5 Proof of Theorem 3 and Corollary 1

Proof. Define

$$\widehat{H}_n := H_n(\hat{\theta}_h), \qquad s_{n,i}(\theta) := \nabla \tilde{l}_{h_n}(\theta; x_i, y_i), \qquad \bar{s}_n := \frac{1}{n}\sum_{i=1}^{n} s_{n,i}(\hat{\theta}_h), \tag{73}$$

$$\widehat{V}_n := \frac{1}{n}\sum_{i=1}^{n}\big(s_{n,i}(\hat{\theta}_h) - \bar{s}_n\big)\big(s_{n,i}(\hat{\theta}_h) - \bar{s}_n\big)^\top, \qquad \widehat{\Sigma}_n := \widehat{H}_n^{-1}\widehat{V}_n\widehat{H}_n^{-1}. \tag{74}$$

Step 1: consistency of $\widehat{H}_n$. Decompose

$$\|\widehat{H}_n - H_0\|_{\mathrm{op}} \le \|H_n(\hat{\theta}_h) - H(\hat{\theta}_h)\|_{\mathrm{op}} + \|H(\hat{\theta}_h) - H(\tilde{\theta})\|_{\mathrm{op}} + \|H(\tilde{\theta}) - H_0\|_{\mathrm{op}}.$$

The first term is $o_p(1)$ by (42). The third is $o_p(1)$ by (60). For the second, note that $\hat{\theta}_h, \tilde{\theta} \in \mathcal{N}_n$ and $\|\hat{\theta}_h - \tilde{\theta}\|_2 = O_p(n^{-1/2}) \to 0$ by (49); the same dominated-convergence argument used in Step 1 of the proof of Theorem 2 therefore gives

$$\|H(\hat{\theta}_h) - H(\tilde{\theta})\|_{\mathrm{op}} = o_p(1).$$

Hence

$$\widehat{H}_n \to H_0. \tag{75}$$

Step 2: consistency of $\widehat{V}_n$. Because $\hat{\theta}_h \in \mathrm{int}(\mathcal{N}_n)$ and $\nabla M_n(\hat{\theta}_h) = 0$ by (51),

$$\bar{s}_n = -\lambda_n J\hat{\theta}_h = o_p(1), \qquad \bar{s}_n\bar{s}_n^\top = o_p(1). \tag{76}$$

It is therefore enough to prove that

$$\frac{1}{n}\sum_{i=1}^{n} s_{n,i}(\hat{\theta}_h)s_{n,i}(\hat{\theta}_h)^\top \to V_0. \tag{77}$$

Write

$$\frac{1}{n}\sum_{i=1}^{n} s_{n,i}(\hat{\theta}_h)s_{n,i}(\hat{\theta}_h)^\top - V_0 = A_{1n} + A_{2n} + A_{3n},$$

where

$$A_{1n} := \frac{1}{n}\sum_{i=1}^{n}\Big[s_{n,i}(\hat{\theta}_h)s_{n,i}(\hat{\theta}_h)^\top - s_{n,i}(\tilde{\theta})s_{n,i}(\tilde{\theta})^\top\Big],$$

$$A_{2n} := \frac{1}{n}\sum_{i=1}^{n}\Big[s_{n,i}(\tilde{\theta})s_{n,i}(\tilde{\theta})^\top - \mathbb{E}\big\{s_{n,1}(\tilde{\theta})s_{n,1}(\tilde{\theta})^\top\big\}\Big],$$

$$A_{3n} := \mathbb{E}\big\{s_{n,1}(\tilde{\theta})s_{n,1}(\tilde{\theta})^\top\big\} - V_0.$$

By (68) and dominated convergence, $A_{3n} \to 0$. By the law of large numbers and the envelope (67), $A_{2n} \to 0$ in probability. For A_{1n}, use the identity

$$aa^\top - bb^\top = (a-b)a^\top + b(a-b)^\top$$

and the mean-value theorem. Because the Hessian satisfies

$$\|\nabla^2 \tilde{l}_{h_n}(\theta; x, y)\|_{\mathrm{op}} \le \frac{C}{h_n}\|x\|_2^2,$$

we have

$$\|s_{n,i}(\hat{\theta}_h) - s_{n,i}(\tilde{\theta})\|_2 \le \frac{C}{h_n}\|x_i\|_2^2\,\|\hat{\theta}_h - \tilde{\theta}\|_2. \tag{78}$$

Together with (46), this gives

$$\|A_{1n}\|_{\mathrm{op}} \le \frac{C}{h_n}\,\|\hat{\theta}_h - \tilde{\theta}\|_2\left(\frac{1}{n}\sum_{i=1}^{n}\|x_i\|_2^3\right).$$

By (49), the law of large numbers, and $\sqrt{n}\,h_n \to \infty$, the right-hand side is $o_p(1)$. Hence (77) holds, and together with (76) we obtain

$$\widehat{V}_n \to V_0. \tag{79}$$

Step 3: sandwich consistency and Wald inference. By (61), $H_0 \succ 0$, so (75) implies that $\widehat{H}_n$ is invertible with probability tending to one. Combining (75) and (79),

$$\widehat{\Sigma}_n = \widehat{H}_n^{-1}\widehat{V}_n\widehat{H}_n^{-1} \to H_0^{-1}V_0H_0^{-1}. \tag{80}$$

This proves Theorem 3.

For Corollary 1, let $a \in \mathbb{R}^{d+1}$ be fixed and nonzero. Theorem 2 and (80) give

$$\frac{\sqrt{n}\,a^\top(\hat{\theta}_h - \theta^*)}{\sqrt{a^\top\widehat{\Sigma}_n a}} \Rightarrow N(0,1)$$

by Slutsky theorem. The usual Wald confidence interval for the jth component follows by taking $a = e_j$. □

A.6 Proofs for Optimal LabelDP Mechanism

We define

$$\omega_i := \frac{(A_i + B_i)^2\|x_i\|_2^2}{n}, \qquad N_i(p_{-1}, p_{+1}) := p_{+1}(1-p_{+1})q_i + p_{-1}(1-p_{-1})(1-q_i),$$

and $\Delta(p_{-1}, p_{+1}) := p_{-1} + p_{+1} - 1$. Then $T(\theta^*; p_{-1}, p_{+1}) = \sum_{i=1}^n \omega_i \frac{N_i(p_{-1}, p_{+1})}{\Delta(p_{-1}, p_{+1})^2}$. Moreover, by the definitions $A_i = 1\{1 - x_i^\top\theta^* > 0\}, B_i = 1\{1 + x_i^\top\theta^* > 0\}$, Hence exactly one of the following three cases occurs:

$$-x_i^\top\theta^* > 1 \Rightarrow (A_i, B_i) = (0,1), -1 < -x_i^\top\theta^* < 1$$

$$\Rightarrow (A_i, B_i) = (1,1), -x_i^\top\theta^* < -1 \Rightarrow (A_i, B_i) = (1,0).$$

At the boundary points $-x_i^\top\theta^* = \pm 1$, one still has $(A_i, B_i) = (0,1)$ or $(1,0)$, respectively. Therefore, $A_i + B_i \in \{1,2\}$, consequently,

$$\omega_i = \frac{(A_i + B_i)^2\|x_i\|_2^2}{n} > 0 \qquad \text{for every } i.$$

Proof of Proposition 2

We define

$$A^\circ := 1\{1 - X^\top\theta^* \geq 0\}, \qquad B^\circ := 1\{1 + X^\top\theta^* \geq 0\}.$$

Then

$$\zeta(\theta^*; X, +1) = -\frac{X}{\Delta}\Big(p_{-1}A^\circ + (1-p_{+1})B^\circ\Big),$$

and

$$\zeta(\theta^*; X, -1) = \frac{X}{\Delta}\Big((1-p_{-1})A^\circ + p_{+1}B^\circ\Big).$$

We apply the law of total variance:

$$V_0 = \mathrm{Var}\Big(E\{\zeta(\theta^*; X, Y) \mid X, Y^*\}\Big) + E\Big[\mathrm{Var}\{\zeta(\theta^*; X, Y) \mid X, Y^*\}\Big].$$

We denote the two terms by

$$V_{\mathrm{np}} := \mathrm{Var}\Big(E\{\zeta(\theta^*; X, Y) \mid X, Y^*\}\Big), \qquad V_{\mathrm{priv}} := E\Big[\mathrm{Var}\{\zeta(\theta^*; X, Y) \mid X, Y^*\}\Big].$$

We first compute the conditional mean.

$$E\{\zeta(\theta^*; X, Y) \mid X, Y^* = +1\} = p_{+1}\zeta(\theta^*; X, +1) + (1 - p_{+1})\zeta(\theta^*; X, -1)$$
$$= \frac{X}{\Delta}\Big[-p_{+1}\{p_{-1}A^\circ + (1 - p_{+1})B^\circ\} + (1 - p_{+1})\{(1 - p_{-1})A^\circ + p_{+1}B^\circ\}\Big].$$

The coefficient of A° is

$$-p_{+1}p_{-1} + (1 - p_{+1})(1 - p_{-1}) = 1 - p_{+1} - p_{-1} = -\Delta,$$

and the coefficient of B° is

$$-p_{+1}(1 - p_{+1}) + (1 - p_{+1})p_{+1} = 0.$$

Hence

$$E\{\zeta(\theta^*; X, Y) \mid X, Y^* = +1\} = -XA^\circ.$$

$$E\{\zeta(\theta^*; X, Y) \mid X, Y^* = -1\} = (1 - p_{-1})\zeta(\theta^*; X, +1) + p_{-1}\zeta(\theta^*; X, -1)$$
$$= \frac{X}{\Delta}\Big[-(1 - p_{-1})\{p_{-1}A^\circ + (1 - p_{+1})B^\circ\} + p_{-1}\{(1 - p_{-1})A^\circ + p_{+1}B^\circ\}\Big].$$

Now the coefficient of A° is

$$-(1 - p_{-1})p_{-1} + p_{-1}(1 - p_{-1}) = 0,$$

while the coefficient of B° is

$$-(1 - p_{-1})(1 - p_{+1}) + p_{-1}p_{+1} = p_{-1} + p_{+1} - 1 = \Delta.$$

Thus

$$E\{\zeta(\theta^*; X, Y) \mid X, Y^* = -1\} = XB^\circ.$$

Combining the two cases yields

$$E\{\zeta(\theta^*; X, Y) \mid X, Y^*\} = -Y^*X\,1\{1 - Y^*X^\top\theta^* \geq 0\}.$$

Therefore,

$$V_{\text{np}} = \text{Var}\Big(-Y^*X\,1\{1 - Y^*X^\top\theta^* \geq 0\}\Big).$$

Next we compute the conditional variance term. If $Y^* = +1$, then

$$\zeta(\theta^*; X, Y) \in \left\{-\frac{X}{\Delta}\big(p_{-1}A^\circ + (1 - p_{+1})B^\circ\big),\ \frac{X}{\Delta}\big((1 - p_{-1})A^\circ + p_{+1}B^\circ\big)\right\},$$

with probabilities p_{+1} and $1 - p_{+1}$, respectively. Hence

$$\text{Var}\{\zeta(\theta^*; X, Y) \mid X, Y^* = +1\} = p_{+1}(1 - p_{+1})(z_+ - z_-)(z_+ - z_-)^\top,$$

where

$$z_+ - z_- = -\frac{X}{\Delta}\Big[p_{-1}A^\circ + (1 - p_{+1})B^\circ + (1 - p_{-1})A^\circ + p_{+1}B^\circ\Big] = -\frac{A^\circ + B^\circ}{\Delta}X.$$

Thus

$$\text{Var}\{\zeta(\theta^*; X, Y) \mid X, Y^* = +1\} = \frac{p_{+1}(1 - p_{+1})}{\Delta^2}(A^\circ + B^\circ)^2XX^\top.$$

Likewise, if $Y^* = -1$, then

$$\text{Var}\{\zeta(\theta^*; X, Y) \mid X, Y^* = -1\} = \frac{p_{-1}(1 - p_{-1})}{\Delta^2}(A^\circ + B^\circ)^2XX^\top.$$

Now let

$$q := P(Y^* = +1 \mid X).$$

Conditioning on X gives

$$\begin{aligned} V_{\mathrm{priv}} &= E\Big[\mathrm{Var}\{\zeta(\theta^*; X, Y) \mid X, Y^*\}\Big] \\ &= E\left[q \cdot \frac{p_{+1}(1-p_{+1})}{\Delta^2}(A^\circ + B^\circ)^2 XX^\top + (1-q) \cdot \frac{p_{-1}(1-p_{-1})}{\Delta^2}(A^\circ + B^\circ)^2 XX^\top\right] \\ &= E\left[\frac{p_{+1}(1-p_{+1})q + p_{-1}(1-p_{-1})(1-q)}{(p_{-1}+p_{+1}-1)^2}(A^\circ + B^\circ)^2 XX^\top\right]. \end{aligned}$$

Finally, by (iv) of Proposition 3,

$$P(1 - X^\top\theta^* = 0) = 0, \qquad P(1 + X^\top\theta^* = 0) = 0,$$

so

$$A^\circ = A := 1\{1 - X^\top\theta^* > 0\}, \qquad B^\circ = B := 1\{1 + X^\top\theta^* > 0\} \quad \text{a.s.}$$

Therefore,

$$V_{\mathrm{np}} = \mathrm{Var}\Big(-Y^* X\, 1\{1 - Y^* X^\top\theta^* \geq 0\}\Big),$$

and

$$V_{\mathrm{priv}} = E\left[\frac{p_{+1}(1-p_{+1})q + p_{-1}(1-p_{-1})(1-q)}{(p_{-1}+p_{+1}-1)^2}(A + B)^2 XX^\top\right].$$

□

Proof of Lemma 1

Differentiate $T(\theta^*; p_{-1}, p_{+1})$ with respect to p_{-1}. Since

$$\frac{\partial}{\partial p_{-1}} N_i(p_{-1}, p_{+1}) = (1-q_i)(1-2p_{-1}), \qquad \frac{\partial}{\partial p_{-1}}\Delta(p_{-1}, p_{+1}) = 1,$$

we obtain

$$\frac{\partial}{\partial p_{-1}} T(\theta^*; p_{-1}, p_{+1}) = \sum_{i=1}^n \omega_i \frac{(1-q_i)(1-2p_{-1})\Delta(p_{-1}, p_{+1}) - 2N_i(p_{-1}, p_{+1})}{\Delta(p_{-1}, p_{+1})^3}.$$

A direct algebraic rearrangement gives

$$\begin{aligned} &(1-q_i)(1-2p_{-1})\Delta(p_{-1}, p_{+1}) - 2N_i(p_{-1}, p_{+1}) \\ &= -(1-q_i)\big\{(1-p_{-1})(1-p_{+1}) + p_{-1}p_{+1}\big\} - 2q_i p_{+1}(1-p_{+1}) < 0. \end{aligned}$$

Hence

$$\frac{\partial}{\partial p_{-1}} T(\theta^*; p_{-1}, p_{+1}) < 0.$$

Similarly, because

$$\frac{\partial}{\partial p_{+1}} N_i(p_{-1}, p_{+1}) = q_i(1-2p_{+1}), \qquad \frac{\partial}{\partial p_{+1}}\Delta(p_{-1}, p_{+1}) = 1,$$

we have

$$\frac{\partial}{\partial p_{+1}} T(\theta^*; p_{-1}, p_{+1}) = \sum_{i=1}^n \omega_i \frac{q_i(1-2p_{+1})\Delta(p_{-1}, p_{+1}) - 2N_i(p_{-1}, p_{+1})}{\Delta(p_{-1}, p_{+1})^3},$$

and again a direct simplification yields

$$\begin{aligned} &q_i(1-2p_{+1})\Delta(p_{-1}, p_{+1}) - 2N_i(p_{-1}, p_{+1}) \\ &= -q_i\big\{(1-p_{-1})(1-p_{+1}) + p_{-1}p_{+1}\big\} - 2(1-q_i)p_{-1}(1-p_{-1}) < 0. \end{aligned}$$

Therefore

$$\frac{\partial}{\partial p_{+1}} T(\theta^*; p_{-1}, p_{+1}) < 0.$$

Consequently, $T(\theta^*; p_{-1}, p_{+1})$ is nonincreasing in each coordinate over the feasible region. Hence there exists a boundary point $(p^*_{-1}, p^*_{+1}) \in \arg\min_{(p_{-1},p_{+1})\in\mathcal{R} \text{ or } \mathcal{R}'} T(\theta^*; p_{-1}, p_{+1})$. □

Proof of Theorem 4

By Lemma 1, it suffices to minimize $T(\theta^*; p_{-1}, p_{+1})$ on the two upper boundary segments of $\mathcal{R}$.

Case 1: the boundary $p_{+1} = e^\varepsilon(1 - p_{-1})$. We write

$$p_{-1} = r(t) := (1-t) + \frac{e^\varepsilon}{e^\varepsilon+1}t = 1 - \frac{t}{e^\varepsilon+1}, \qquad p_{+1} = s(t) := \frac{e^\varepsilon}{e^\varepsilon+1}t, \qquad t \in (0,1].$$

Then

$$r(t) + s(t) - 1 = \frac{e^\varepsilon - 1}{e^\varepsilon + 1}t > 0,$$

and

$$r(t)\{1-r(t)\} = \frac{t(e^\varepsilon+1-t)}{(e^\varepsilon+1)^2}, \qquad s(t)\{1-s(t)\} = \frac{e^\varepsilon t(e^\varepsilon+1-e^\varepsilon t)}{(e^\varepsilon+1)^2}.$$

Substituting these identities into $T(\theta^*; p_{-1}, p_{+1})$, we get

$$\begin{aligned} T\big(\theta^*; r(t), s(t)\big) &= \sum_{i=1}^n \omega_i \frac{s(t)\{1-s(t)\}q_i + r(t)\{1-r(t)\}(1-q_i)}{\{r(t)+s(t)-1\}^2} \\ &= \frac{1}{(e^\varepsilon-1)^2}\sum_{i=1}^n \omega_i \left[\frac{(e^\varepsilon+1)\{1+(e^\varepsilon-1)q_i\}}{t} - \{1+(e^{2\varepsilon}-1)q_i\}\right]. \end{aligned}$$

Therefore,

$$\frac{d}{dt}T\big(\theta^*; r(t), s(t)\big) = -\frac{e^\varepsilon+1}{(e^\varepsilon-1)^2t^2}\sum_{i=1}^n \omega_i\{1+(e^\varepsilon-1)q_i\} < 0.$$

Hence $T\big(\theta^*; r(t), s(t)\big)$ is monotonically decreasing in t, so its minimum on this boundary segment is attained at $t = 1$. At $t = 1$,

$$p_{-1} = r(1) = \frac{e^\varepsilon}{e^\varepsilon+1}, \qquad p_{+1} = s(1) = \frac{e^\varepsilon}{e^\varepsilon+1}.$$

Case 2: the boundary $p_{-1} = e^\varepsilon(1 - p_{+1})$. Interchanging the roles of p_{-1} and p_{+1} gives the analogous parameterization

$$p_{-1} = s(t), \qquad p_{+1} = r(t), \qquad t \in (0,1],$$

and the same calculation shows that the minimum is again achieved at $t = 1$. Thus the minimizing point on the second boundary segment is also

$$\left(\frac{e^\varepsilon}{e^\varepsilon+1}, \frac{e^\varepsilon}{e^\varepsilon+1}\right).$$

Combining the two cases, the global minimizer over $\mathcal{R}$ is

$$p_{-1} = p_{+1} = \frac{e^\varepsilon}{e^\varepsilon+1}.$$

□

Proof of Theorem 5

Let

$$\alpha_{\varepsilon,\delta} := \frac{e^\varepsilon+\delta}{e^\varepsilon+1}.$$

By Lemma 1, it is enough to minimize $T(\theta^*; p_{-1}, p_{+1})$ on the two upper boundary segments of $\mathcal{R}'$.

Case 1: the boundary $p_{+1} = e^\varepsilon(1 - p_{-1}) + \delta$. We write

$$p_{-1} = r_\delta(t) := \big(1 + e^{-\varepsilon}\delta\big)(1-t) + \frac{e^\varepsilon+\delta}{e^\varepsilon+1}t, \qquad p_{+1} = s_\delta(t) := \frac{e^\varepsilon+\delta}{e^\varepsilon+1}t,$$

for $t \in [t_0, 1]$, where

$$t_0 := \frac{\delta(e^\varepsilon + 1)}{e^\varepsilon + \delta}.$$

Then

$$\big(r_\delta(t_0), s_\delta(t_0)\big) = (1, \delta), \qquad \big(r_\delta(1), s_\delta(1)\big) = \big(\alpha_{\varepsilon,\delta}, \alpha_{\varepsilon,\delta}\big).$$

Equivalently, if we use

$$u := s_\delta(t) \in [\delta, \alpha_{\varepsilon,\delta}],$$

then

$$p_{+1} = u, \qquad p_{-1} = 1 - e^{-\varepsilon}(u - \delta).$$

Define

$$\tau(u) := T\big(\theta^*; 1 - e^{-\varepsilon}(u - \delta), u\big), \qquad u \in [\delta, \alpha_{\varepsilon,\delta}].$$

Since

$$1 - p_{-1} = e^{-\varepsilon}(u - \delta), \qquad p_{-1}(1 - p_{-1}) = e^{-2\varepsilon}(u - \delta)(e^\varepsilon - u + \delta),$$

and

$$p_{-1} + p_{+1} - 1 = e^{-\varepsilon}\big\{(e^\varepsilon - 1)u + \delta\big\},$$

substituting into $T(\theta^*; p_{-1}, p_{+1})$ yields

$$\tau(u) = \sum_{i=1}^{n} \omega_i \frac{e^{2\varepsilon} q_i u(1-u) + (1 - q_i)(u - \delta)(e^\varepsilon - u + \delta)}{\big\{(e^\varepsilon - 1)u + \delta\big\}^2}.$$

Differentiating with respect to u and collecting like terms, we obtain

$$\tau'(u) = -\frac{e^\varepsilon}{\big\{(e^\varepsilon - 1)u + \delta\big\}^3}\big(C_0 + C_1 u\big),$$

where

$$C_0 := \sum_{i=1}^{n} \omega_i\, \delta\Big[(e^\varepsilon - 1 + 2\delta)q_i - (2e^\varepsilon + 2\delta - 1)\Big],$$

and

$$C_1 := (2\delta + e^\varepsilon - 1) \sum_{i=1}^{n} \omega_i \Big\{1 + (e^\varepsilon - 1)q_i\Big\}.$$

Because $e^\varepsilon > 1$, $\delta \in (0, 1)$, $\omega_i > 0$ for every i, and

$$1 + (e^\varepsilon - 1)q_i > 0,$$

we obtain $C_1 > 0$. Also,

$$\frac{e^\varepsilon}{\big\{(e^\varepsilon - 1)u + \delta\big\}^3} > 0 \qquad \text{for all } u \in [\delta, \alpha_{\varepsilon,\delta}].$$

Hence $\tau'(u)$ is a negative multiple of the affine function $C_0 + C_1 u$, namely

$$\tau'(u) = -\frac{e^\varepsilon}{\big\{(e^\varepsilon - 1)u + \delta\big\}^3}(C_0 + C_1 u), \qquad C_1 > 0.$$

Therefore the sign of $\tau'(u)$ is governed by a linear function with positive slope before multiplication by a strictly negative coefficient, so $\tau'(u)$ itself is strictly decreasing in sign pattern and can change sign at most once, necessarily from positive to negative. In particular, τ cannot have an interior local minimum on $[\delta, \alpha_{\varepsilon,\delta}]$; any interior critical point can only be a local maximum. Consequently, the minimum of $\tau(u)$ must occur at one of the two endpoints. Therefore the only candidates on this boundary are

$$(1, \delta) \qquad \text{and} \qquad \big(\alpha_{\varepsilon,\delta}, \alpha_{\varepsilon,\delta}\big).$$

Case 2: the boundary $p_{-1} = e^\varepsilon(1 - p_{+1}) + \delta$. Repeating the same argument with the roles of p_{-1} and p_{+1} interchanged shows that the minimum on this boundary segment is attained at one of its two endpoints, namely

$$(\delta, 1) \qquad \text{or} \qquad \big(\alpha_{\varepsilon,\delta}, \alpha_{\varepsilon,\delta}\big).$$

Combining the two boundary analyses, every global minimizer over $\mathcal{R}'$ must belong to the three-point set

$$\left\{\left(\alpha_{\varepsilon,\delta}, \alpha_{\varepsilon,\delta}\right), (1,\delta), (\delta,1)\right\}.$$

Therefore the optimal (ε,δ)-LabelDP randomized response mechanism is obtained by comparing the three values of T:

(i) If $T(\theta^*; \frac{e^\varepsilon+\delta}{e^\varepsilon+1}, \frac{e^\varepsilon+\delta}{e^\varepsilon+1}) \le T(\theta^*; 1, \delta)$ *and* $T(\theta^*; \frac{e^\varepsilon+\delta}{e^\varepsilon+1}, \frac{e^\varepsilon+\delta}{e^\varepsilon+1}) \le T(\theta^*; \delta, 1)$*, then the optimal randomized response mechanism is given by the probabilities* $p_{-1} = p_{+1} = \frac{e^\varepsilon+\delta}{e^\varepsilon+1}$.

(ii) If $T(\theta^*; 1, \delta) \le T(\theta^*; \frac{e^\varepsilon+\delta}{e^\varepsilon+1}, \frac{e^\varepsilon+\delta}{e^\varepsilon+1})$ *and* $T(\theta^*; 1, \delta) \le T(\theta^*; \delta, 1)$*, then the optimal randomized response mechanism is given by the probabilities* $p_{-1} = 1$ and $p_{+1} = \delta$.

(iii) If $T(\theta^*; \delta, 1) \le T(\theta^*; \frac{e^\varepsilon+\delta}{e^\varepsilon+1}, \frac{e^\varepsilon+\delta}{e^\varepsilon+1})$ *and* $T(\theta^*; \delta, 1) \le T(\theta^*; 1, \delta)$*, then the optimal randomized response mechanism is given by the probabilities* $p_{-1} = \delta$ and $p_{+1} = 1$.

□