The Score-Matched Optimal Convex Estimator Does Not Attain the Full Semiparametric Efficiency Bound

Abstract

We clarify the semiparametric efficiency status of the score-matched optimal convex M-estimator introduced by Feng et al. [2026]. Their estimator minimizes asymptotic variance over convex loss functions in the linear model $Y = X^\top \beta_0 + \varepsilon$ with unknown error density p_0. The optimal convex estimator has asymptotic covariance equal to the semiparametric information bound if and only if p_0 is log-concave. For non-log-concave p_0, its covariance exceeds the bound by $i(p_0) - i^\star(p_0)$; moreover, in an explicit bimodal-mixture/Rademacher model, the kernel-density-based regression estimator (KDRE) of Yao and Zhao [2013] achieves the smaller variance $1/i(p_0)$. We derive the exact Pythagorean identity $i(p_0) - i^\star(p_0) = \mathrm{dist}_{L_2(P_0)}(\psi_0, \Psi_\downarrow(p_0))^2$, characterizing the efficiency gap as the squared L_2-distance from the true score to the cone of non-increasing functions. We also note that in the unrestricted (nonsymmetric) no-intercept model, the correct semiparametric benchmark involves $\mathrm{Var}(X)$ rather than $\mathbb{E}[XX^\top]$, a correction that is relevant to the problem formulation on SolveAll.org.

1 Introduction

Let $p \in \mathbb{N}$ be fixed. Consider the linear regression model

$$Y_i = X_i^\top \beta_0 + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where $(X_i, Y_i)_{i=1}^n$ are i.i.d., $\beta_0 \in \mathbb{R}^p$, $X_i \in \mathbb{R}^p$, $\varepsilon_i \perp X_i$, and the conditional density of Y_i given $X_i = x$ is $y \mapsto p_0(y - x^\top \beta_0)$ for an unknown error density p_0. This is a classical semiparametric model with Euclidean parameter β_0 and infinite-dimensional nuisance p_0.

A convex M-estimator minimizes $\sum_{i=1}^n \rho(Y_i - X_i^\top \beta)$ for a convex loss ρ. Convexity of ρ is equivalent to monotone non-decreasingness of ρ'. The first-order condition $\sum_i \rho'(\hat\varepsilon_i)\, X_i = 0$ can be written as $\sum_i \psi(\hat\varepsilon_i)\, X_i = 0$ with $\psi = -\rho'$, which is non-increasing for convex ρ. Feng et al. [2026] work with this non-increasing score ψ. (Note that the sign convention $\psi = -\rho'$ follows Feng et al. [2026] and differs from the more common robust-statistics convention $\psi = \rho'$; see Huber and Ronchetti [2009].) The key observation is that the location score $\psi_0 := p_0'/p_0$ is itself non-increasing precisely when p_0 is log-concave. Writing $\Psi_\downarrow(p_0)$ for the cone of non-increasing $L_2(P_0)$-functions, the asymptotic variance of a regular convex M-estimator $\hat\beta_\psi$ based on score ψ takes the form

$$\sqrt{n}\,(\hat\beta_\psi - \beta_0) \Rightarrow N\big(0,\, V_{p_0}(\psi)\, \Sigma_X^{-1}\big), \qquad \Sigma_X := \mathbb{E}[XX^\top], \tag{2}$$

where $V_{p_0}(\psi) := \int \psi^2 \, dP_0 \big/ \big(\int_{S_0} p_0 \, d\psi\big)^2$ and $S_0 = \{z : p_0(z) > 0\}$.

Feng et al. [2026] prove that the population optimizer of the score-matching objective $D_{p_0}(\psi) := \int \psi^2 \, dP_0 + 2 \int_{S_0} p_0 \, d\psi$ over $\Psi_\downarrow(p_0)$ is $\psi_0^\star := \widehat{J}_0^{(R)} \circ F_0$, where $\widehat{J}_0$ is the least concave majorant of the density-quantile function $J_0 := p_0 \circ F_0^{-1}$ and $\widehat{J}_0^{(R)}$ is its right derivative. Their Theorem 2 establishes that the corresponding "antitonic Fisher information" satisfies

$$0 < i^\star(p_0) := \int (\psi_0^\star)^2 \, dP_0 \;\leq\; i(p_0) := \int \psi_0^2 \, dP_0, \tag{3}$$

with equality if and only if p_0 is log-concave, where $\psi_0 := p_0'/p_0$ is the location score.

Their sample-level score-matched convex estimator $\hat{\beta}_{\text{cvx}}$ is constructed via a five-step procedure: (1) sample splitting into three folds, (2) pilot estimation and residual computation, (3) kernel density estimation of the error density on each fold, (4) antitonic projection of the preliminary score onto the monotone cone via the Pool Adjacent Violators Algorithm (PAVA), and (5) cross-fitted convex M-estimation. In the symmetric no-intercept model, Theorem 13 of Feng et al. [2026] gives

$$\sqrt{n}(\hat{\beta}_n^{\text{sym}} - \beta_0) \xrightarrow{d} N\big(0, \{\mathbb{E}(XX^\top)\}^{-1}/i^\star(p_0)\big). \tag{4}$$

In the intercept model $Y_i = \mu_0 + \tilde{X}_i^\top \theta_0 + \varepsilon_i$ with a location-identifying centering condition $\mathbb{E}\zeta(\varepsilon_1) = 0$, Theorem 14 gives the analogous result with $\mathrm{Cov}(\tilde{X})^{-1}/i^\star(p_0)$ for the slope. Meanwhile, Proposition 38(a) computes the full semiparametric efficient score as $\tilde{\ell}_{\theta_0}(x, y) = -(\tilde{x} - \tilde{m})\psi_0(y - x^\top \beta_0)$ with efficient information $\tilde{I}_{\theta_0} = i(p_0)\,\Sigma$, where $\Sigma = \mathrm{Cov}(\tilde{X})$ and $\tilde{m} = \mathbb{E}[\tilde{X}_1]$. Feng et al. explicitly describe Theorem 14 as an "antitonic efficiency" result: the convex estimator's covariance is identical to the semiparametric efficient bound, except with $i^\star(p_0)$ in place of $i(p_0)$.

A natural question, raised by Feng et al. [2026] and posed as an open problem on SolveAll.org [SolveAll.org, 2026b], is whether the convex-class optimum already coincides with the full semiparametric efficiency bound for unknown p_0.

We show that the convex estimator matches the semiparametric lower bound if and only if p_0 is log-concave. Furthermore, in at least some non-log-concave models—including our explicit Gaussian-mixture/Rademacher example—a strictly better nonconvex regular estimator exists. Whether the full bound is attainable by some regular estimator under only finite Fisher information (without the smoothness conditions required by the KDRE) remains open; see Section 6.

1.1 Contributions

This short note combines standard semiparametric tangent-space calculations with the projection results already established by Feng et al. [2026]. The genuinely distinctive contributions are:

(i) In the unrestricted (nonsymmetric) no-intercept model, we show that the correct semiparametric benchmark involves $\mathrm{Var}(X)$ rather than $\mathbb{E}[XX^\top]$, a correction relevant to the current problem formulation on SolveAll.org (Theorem 2).

(ii) We construct an explicit counterexample—Rademacher design with bimodal Gaussian mixture errors—in which the KDRE of Yao and Zhao [2013] achieves strictly smaller asymptotic variance than the convex optimum (Theorem 3).

We also include, for completeness, the classical efficiency bound in the symmetric model (Theorem 1) and a Pythagorean identity characterizing the efficiency gap geometrically (Theorem 4); these are clean consequences of standard projection theory and the cone-projection result in Feng et al. [2026].

1.2 Related work

Adaptive estimation in the linear model with unknown error density has a long history. Starting from the seminal work of Stein [1956], a series of contributions [van Eeden, 1970, Stone, 1975, Beran, 1978, Bickel, 1982, Schick, 1986] established that adaptive, asymptotically efficient estimators of location and regression parameters can be constructed without knowledge of the error density. Portnoy and Koenker [1989] constructed asymptotically efficient adaptive L-estimators for slope parameters under finite Fisher information. Yao and Zhao [2013] proved that a kernel-density-based regression estimator (KDRE) achieves oracle MLE efficiency in the exact model $Y = X^\top\beta + \varepsilon$ with $X \perp \varepsilon$.

The novelty of Feng et al. [2026], now published in *The Annals of Statistics*, is to characterize the best *convex* surrogate to the efficient score, yielding computational stability and robustness guarantees at the cost of a potential efficiency loss. Their approach connects score matching [Hyvärinen, 2005] with monotone function estimation [Groeneboom and Jongbloed, 2014] and classical robust statistics [Huber, 1964, Huber and Ronchetti, 2009]. Notably, for Cauchy errors the optimal convex loss turns out to be Huber-like, and for heavy-tailed errors generally the projected score is bounded, ensuring robustness in the sense of finite gross error sensitivity [Hampel, 1974]. Their R package `asm` implements the full procedure.

2 Semiparametric Efficiency in the Symmetric No-Intercept Model

The semiparametric efficiency bounds in this and the next section are classical, following from the general tangent-space projection theory of Bickel [1982] and van der Vaart [1998]. We include the derivations to make explicit how the resulting efficient information $i(p_0)$ compares with the antitonic information $i^\star(p_0)$ of Feng et al. [2026].

Consider the semiparametric model

$$\mathcal{P}_{\mathrm{sym}} = \big\{P_{\beta,Q,p} : (X,Y) \sim Q(dx)\, p(y - x^\top\beta)\, dy,\ p(z) = p(-z)\big\}, \tag{5}$$

at a true law $P_0 = P_{\beta_0,Q_0,p_0}$ with p_0 absolutely continuous, $i(p_0) < \infty$, and $\mathbb{E}_{Q_0}[XX^\top]$ positive definite.

Theorem 1 (Symmetric no-intercept model)**.** *In the model $\mathcal{P}_{\mathrm{sym}}$, the efficient score for β_0 in direction $h \in \mathbb{R}^p$ is*

$$\ell_h^{\mathrm{eff}}(X,Y) = -h^\top X\, \psi_0(\varepsilon),$$

the efficient information is $I_{\mathrm{eff}} = i(p_0)\,\mathbb{E}[XX^\top]$, and the semiparametric efficiency bound is

$$I_{\mathrm{eff}}^{-1} = \frac{\mathbb{E}[XX^\top]^{-1}}{i(p_0)}.$$

Consequently, the score-matched optimal convex estimator of Feng et al. [2026] attains this bound if and only if p_0 is log-concave.

Proof. The proof proceeds in four steps.

Step 1: Nuisance tangent space. Let $L_2^0(Q_0)$ denote square-integrable mean-zero functions of X, and let

$$L_{2,\mathrm{even}}^0(P_0) := \big\{\Gamma \in L_2(P_0) : \mathbb{E}[\Gamma(\varepsilon)] = 0,\ \Gamma(z) = \Gamma(-z) \text{ a.e.}\big\}.$$

The nuisance tangent space is

$$\mathcal{N}_{\mathrm{sym}} = \big\{a(X) + \Gamma(\varepsilon) : a \in L_2^0(Q_0),\ \Gamma \in L_{2,\mathrm{even}}^0(P_0)\big\}.$$

This is the closed direct sum of two orthogonal L_2-subspaces, generated by regular parametric submodels: for any bounded a with $\mathbb{E}[a(X)] = 0$, the tilted design $dQ_t/dQ_0 = 1 + ta$ is valid for small $|t|$; for any bounded even Γ with $\mathbb{E}[\Gamma(\varepsilon)] = 0$, the tilted density $p_t(z) = p_0(z)(1 + t\Gamma(z))$ is a valid symmetric density for small $|t|$.

Step 2: Parametric score. For direction $h \in \mathbb{R}^p$, the log-likelihood score along $\beta_t = \beta_0 + th$ is

$$S_h(X,Y) = -h^\top X\, \psi_0(\varepsilon), \qquad \varepsilon := Y - X^\top \beta_0,$$

since $\frac{\partial}{\partial t}\big|_{t=0} \log p_0(Y - X^\top(\beta_0 + th)) = \frac{p_0'(\varepsilon)}{p_0(\varepsilon)} \cdot (-h^\top X) = -h^\top X\, \psi_0(\varepsilon)$.

Step 3: Orthogonality. The two summands of $\mathcal{N}_{\mathrm{sym}}$ are orthogonal in $L_2(P_0)$ because $X \perp \varepsilon$ and each summand has mean zero: for $a \in L_2^0(Q_0)$ and $\Gamma \in L^0_{2,\mathrm{even}}(P_0)$, $\mathbb{E}[a(X)\Gamma(\varepsilon)] = \mathbb{E}[a(X)]\mathbb{E}[\Gamma(\varepsilon)] = 0$. Since both subspaces are closed (the first depends only on X, the second only on ε), $\mathcal{N}_{\mathrm{sym}}$ is a closed subspace.

For $a \in L_2^0(Q_0)$: by independence of X and ε,

$$\mathbb{E}[S_h \cdot a(X)] = -h^\top \mathbb{E}[X a(X)]\, \mathbb{E}[\psi_0(\varepsilon)] = 0,$$

since $\mathbb{E}[\psi_0(\varepsilon)] = \int p_0'(z)\, dz = 0$.

For even $\Gamma \in L^0_{2,\mathrm{even}}(P_0)$: since p_0 is symmetric, $\psi_0 = p_0'/p_0$ is odd, so the product $\psi_0 \cdot \Gamma$ is odd, and therefore

$$\mathbb{E}[\psi_0(\varepsilon)\Gamma(\varepsilon)] = \int \psi_0(z)\Gamma(z)\, p_0(z)\, dz = 0.$$

By independence, $\mathbb{E}[S_h \cdot \Gamma(\varepsilon)] = -h^\top \mathbb{E}[X]\, \mathbb{E}[\psi_0(\varepsilon)\Gamma(\varepsilon)] = 0$. Therefore $S_h \perp \mathcal{N}_{\mathrm{sym}}$.

Step 4: Conclusion. By the orthogonal projection theorem in $L_2(P_0)$, since $S_h \in \mathcal{N}_{\mathrm{sym}}^\perp$, the efficient score is $S_h^{\mathrm{eff}} = S_h = -h^\top X\, \psi_0(\varepsilon)$, giving efficient information $I_{\mathrm{eff}} = i(p_0)\, \mathbb{E}[XX^\top]$.

Comparing with the convex estimator covariance $\mathbb{E}[XX^\top]^{-1}/i^\star(p_0)$ from Theorem 13 of Feng et al. [2026], and using Theorem 2(d) therein ($i^\star(p_0) = i(p_0)$ iff p_0 is log-concave), the result follows. $\square$

3 The Unrestricted No-Intercept Model

We now drop the symmetry assumption. The resulting tangent-space calculation is again standard [van der Vaart, 1998], but leads to a centering correction that changes the semiparametric benchmark from $\mathbb{E}[XX^\top]$ to $\mathrm{Var}(X)$.

Consider

$$\mathcal{P} = \left\{ P_{\beta,Q,p} : (X,Y) \sim Q(dx)\, p(y - x^\top \beta)\, dy \right\}, \tag{6}$$

with p_0 absolutely continuous, $i(p_0) < \infty$, and $\mathrm{Var}(X)$ positive definite. We require $\mathrm{Var}(X) \succ 0$ rather than $\mathbb{E}[XX^\top] \succ 0$ because, once the error density p_0 is unrestricted (no symmetry or mean-zero constraint), the nuisance tangent space contains *all* mean-zero functions of ε. In particular, the score $-h^\top X\, \psi_0(\varepsilon)$ is no longer orthogonal to the nuisance directions in ε: projecting onto $L_2^0(P_0)$ removes the $\mathbb{E}[X]$-component, leaving $-(X - \mathbb{E}[X])\, \psi_0(\varepsilon)$. The identifiable component of β_0 is therefore determined by the centered design $X - \mathbb{E}[X]$, and the relevant design matrix is $\mathrm{Var}(X)$, not $\mathbb{E}[XX^\top]$.

Theorem 2 (Unrestricted no-intercept model)**.** *In the model $\mathcal{P}$, the efficient score for β_0 in direction h is*

$$\ell_h^{\mathrm{eff}}(X,Y) = -h^\top \big(X - \mathbb{E}[X]\big)\, \psi_0(\varepsilon),$$

the efficient information is $I_{\mathrm{eff}} = i(p_0)\,\mathrm{Var}(X)$, and the semiparametric efficiency bound is

$$I_{\mathrm{eff}}^{-1} = \frac{\mathrm{Var}(X)^{-1}}{i(p_0)}.$$

In particular, the SolveAll benchmark using $\mathbb{E}[XX^\top]$ is correct only when $\mathbb{E}[X] = 0$.

Proof. The nuisance tangent space is now

$$\mathcal{N} = \{a(X) + \Gamma(\varepsilon) : a \in L_2^0(Q_0),\ \Gamma \in L_2^0(P_0)\},$$

without the even-symmetry restriction on Γ. As before, $X \perp \varepsilon$ and the mean-zero conditions ensure the two summands are orthogonal and each generates a closed subspace of $L_2(P_0)$, so $\mathcal{N}$ is closed.

The score $S_h = -h^\top X\,\psi_0(\varepsilon)$ is still orthogonal to $L_2^0(Q_0)$ by the argument of Theorem 1. However, it is no longer orthogonal to $L_2^0(P_0)$ in general: the projection of S_h onto $L_2^0(P_0)$ is obtained via the conditional expectation

$$\mathbb{E}[S_h \mid \varepsilon] = -h^\top \mathbb{E}[X]\,\psi_0(\varepsilon),$$

which lies in $L_2^0(P_0)$ since $\mathbb{E}[\psi_0(\varepsilon)] = 0$. (When p_0 is symmetric, this projection vanishes because ψ_0 is odd and gets absorbed by the even-symmetry restriction.) The efficient score is the residual after projecting S_h onto $\mathcal{N}$:

$$S_h^{\mathrm{eff}} = S_h - \mathbb{E}[S_h \mid \varepsilon] = -h^\top (X - \mathbb{E}[X])\,\psi_0(\varepsilon),$$

with efficient information $\mathbb{E}[(S_h^{\mathrm{eff}})^2]/(h^\top h) = i(p_0)\,\mathrm{Var}(X)$, where we used $\mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] = \mathrm{Var}(X)$ and $\mathbb{E}[\psi_0(\varepsilon)^2] = i(p_0)$.

Sanity check. If one adjoins an intercept by writing $X = (\tilde{X}, 1)$, then $\mathrm{Var}(X)$ becomes singular, reflecting the non-identifiability of the intercept without a centering restriction. Restricting to slopes recovers Proposition 38(a) of Feng et al. [2026]: the efficient score for the slope is $-(\tilde{X} - \mathbb{E}[\tilde{X}])\,\psi_0(\varepsilon)$ with efficient information $i(p_0)\,\mathrm{Cov}(\tilde{X})$. $\square$

Remark 1. *In Feng et al. [2026]'s intercept/location-identified model $\mathcal{Q}_\zeta$, Theorem 14 gives the convex slope covariance $\mathrm{Cov}(\tilde{X})^{-1}/i^\star(p_0)$, while Proposition 38(a) gives the semiparametric bound $\mathrm{Cov}(\tilde{X})^{-1}/i(p_0)$. The comparison again reduces to $i^\star(p_0)$ versus $i(p_0)$, with equality iff p_0 is log-concave.*

4 A Concrete Counterexample

Theorem 3 (Counterexample)**.** *There exists a valid submodel of the SolveAll setup in which the score-matched optimal convex estimator has strictly larger asymptotic variance than the full semiparametric efficiency bound. In particular, a nonconvex regular estimator achieves strictly smaller asymptotic variance than the convex optimum.*

Proof. Take $p = 1$. Let

$$X_i \in \{-1, 1\}, \quad P(X_i = 1) = P(X_i = -1) = \tfrac{1}{2},$$

and let the errors be independent of the design with density

$$p_a(z) := \tfrac{1}{2}\,\phi(z - a) + \tfrac{1}{2}\,\phi(z + a), \qquad a > 1, \tag{7}$$

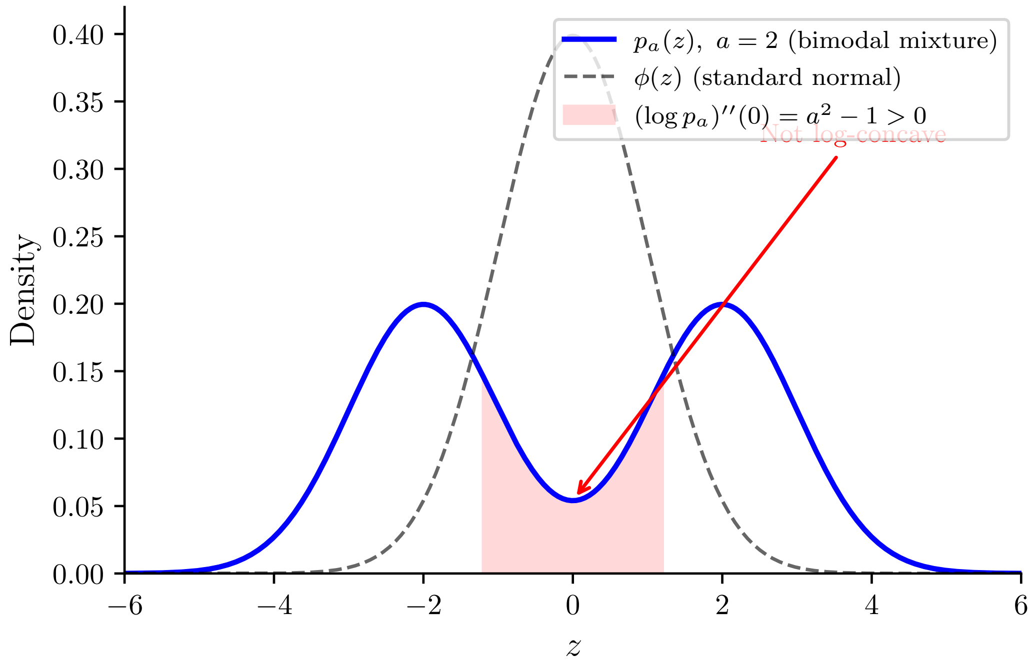

Figure 1: Bimodal Gaussian mixture error density $p_a(z) = \frac{1}{2}\phi(z-2) + \frac{1}{2}\phi(z+2)$ (solid blue) versus the standard normal density $\phi(z)$ (dashed). The mixture is symmetric but not log-concave: at $z = 0$, $(\log p_a)''(0) = a^2 - 1 = 3 > 0$ (shaded region), so the log-density is locally convex rather than concave. By Theorem 2(d) of Feng et al. [2026], this non-log-concavity implies a strict gap $i^\star(p_a) < i(p_a)$, and thus the score-matched optimal convex estimator does not attain the semiparametric efficiency bound for this error density.

where ϕ denotes the standard normal density. The model $Y_i = \beta_0 X_i + \varepsilon_i$ is exactly of the SolveAll form with $X \perp \varepsilon$.

Step 1: Non-log-concavity. We compute

$$p_a'(z) = -\tfrac{1}{2}(z-a)\phi(z-a) - \tfrac{1}{2}(z+a)\phi(z+a),$$

so $p_a'(0) = 0$ and

$$(\log p_a)''(0) = \frac{p_a''(0)}{p_a(0)} = a^2 - 1 > 0$$

for $a > 1$. Hence p_a is not log-concave.

Step 2: Strict gap in information. Since p_a is absolutely continuous with finite Fisher information ($|\psi_a(z)| \le |z| + a$ and p_a has Gaussian tails), Theorem 2(d) of Feng et al. [2026] applies. As p_a is not log-concave, we obtain

$$i^\star(p_a) < i(p_a).$$

Theorem 13 gives the optimal convex covariance $1/i^\star(p_a)$ (since $\mathbb{E}[X^2] = 1$).

Step 3: Semiparametric bound. Since p_a is symmetric and $\mathbb{E}[X] = 0$, Theorem 1 gives the full semiparametric efficiency bound $1/i(p_a)$.

Step 4: Attainability via KDRE. Theorem 2.1 of Yao and Zhao [2013] states that the KDRE achieves asymptotic variance $\mathbb{E}[X^2]^{-1}/i(p_0)$ in the model $Y = X^\top\beta + \varepsilon$ with $X \perp \varepsilon$, provided Conditions C1–C5 hold. We verify these for our setup:

- **C1** (design and error): (X_i, ε_i) are i.i.d. and independent; $\mathbb{E}[\varepsilon_i] = 0$ (by symmetry of p_a), $\mathbb{E}|\varepsilon_i|^3 < \infty$ (Gaussian tails), and X_i has bounded support ($\{-1, 1\}$).
- **C2** (error density smoothness): p_a is symmetric with bounded continuous derivatives of all orders (Gaussian mixture).
- **C3** (kernel): Choose any symmetric compactly supported C^4 kernel (e.g., a degree-6 polynomial kernel).
- **C4** (bandwidth): Choose $h_n = n^{-1/6}$, giving $nh_n^4 \to \infty$ and $nh_n^8 \to 0$.
- **C5** (initial estimator): The ordinary least-squares estimator serves as the required initial $\sqrt{n}$-consistent estimator of β_0.

In the notation of Yao and Zhao [2013], their $I_{\beta_0} = \int (\psi_0)^2 \, dP_0 = i(p_a)$ and $M = \mathbb{E}[XX^\top] = \mathbb{E}[X^2] = 1$. Theorem 2.1 therefore yields

$$\sqrt{n}(\hat{\beta}_{\mathrm{KDRE}} - \beta_0) \Rightarrow N\Big(0, \frac{M^{-1}}{I_{\beta_0}}\Big) = N\Big(0, \frac{1}{i(p_a)}\Big).$$

Step 5: Comparison. Since $i^\star(p_a) < i(p_a)$,

$$\mathrm{Avar}(\hat{\beta}_{\mathrm{cvx}}) = \frac{1}{i^\star(p_a)} > \frac{1}{i(p_a)} = \mathrm{Avar}(\hat{\beta}_{\mathrm{KDRE}}). \qquad \square$$

Remark 2 (Cauchy errors)**.** *For the standard Cauchy density $p_0(z) = 1/(\pi(1+z^2))$, which has score $\psi_0(z) = -2z/(1+z^2)$ and Fisher information $i(p_0) = 1/2$, Feng et al. [2026] derive in their Example 12 a closed-form expression for the projected score $\psi_0^\star$. The optimal convex loss $\ell_0^\star = -\log p_0^\star$ is approximately quadratic on a central interval $[-z_0, z_0]$ and linear outside it–resembling a Huber loss with automatically chosen transition point. The resulting antitonic information is $i^\star(p_0) \approx 0.439$, giving a relative efficiency $\mathrm{ARE}^\star(p_0) = i^\star(p_0)/i(p_0) \approx 0.878$. Thus the convex restriction loses approximately 12.2% of the available score information, a modest price for the computational and robustness advantages of convexity.*

5 The Pythagorean Identity for the Efficiency Gap

Theorem 4 (Gap identity)**.** *Let p_0 be absolutely continuous with $i(p_0) < \infty$. Then*

$$i(p_0) - i^\star(p_0) = \mathrm{dist}_{L_2(P_0)}\big(\psi_0, \, \Psi_\downarrow(p_0)\big)^2. \tag{8}$$

Proof. By Theorem 2 of Feng et al. [2026], $\psi_0^\star$ is the $L_2(P_0)$-projection of ψ_0 onto the closed convex cone $\Psi_\downarrow(p_0)$. Since $0 \in \Psi_\downarrow(p_0)$ and $2\psi_0^\star \in \Psi_\downarrow(p_0)$ (the cone is closed under non-negative scaling), the projection optimality condition on the cone gives

$$\langle \psi_0 - \psi_0^\star, \, \psi_0^\star \rangle_{L_2(P_0)} = 0.$$

This is the Moreau decomposition / Pythagorean identity on a convex cone. Therefore

$$\begin{aligned} i(p_0) = \|\psi_0\|_{L_2(P_0)}^2 &= \|\psi_0 - \psi_0^\star\|_{L_2(P_0)}^2 + \|\psi_0^\star\|_{L_2(P_0)}^2 \\ &= \mathrm{dist}_{L_2(P_0)}(\psi_0, \Psi_\downarrow(p_0))^2 + i^\star(p_0). \end{aligned} \qquad \square$$

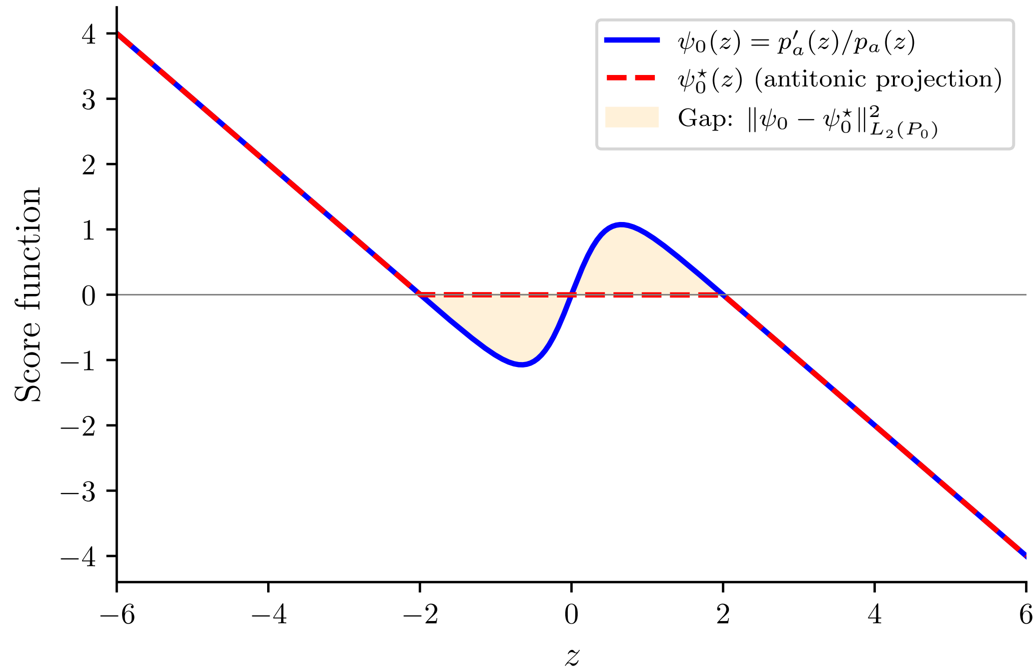

Figure 2: True location score $\psi_0(z) = p_a'(z)/p_a(z)$ (solid blue) and its antitonic projection $\psi_0^\star(z)$ (dashed red) for the bimodal Gaussian mixture with $a = 2$. The true score is non-monotone between the two modes: it increases in the inter-modal region where the density has a local minimum. The projection $\psi_0^\star$, obtained via the least concave majorant construction of Feng et al. [2026], is constrained to be non-increasing and flattens through the non-monotone region. The shaded area between the curves, weighted by p_a, equals the squared $L_2(P_0)$-distance $\|\psi_0 - \psi_0^\star\|^2_{L_2(P_0)} = i(p_a) - i^\star(p_a) \approx 0.226$, which is the efficiency gap of the convex restriction. Numerically, $i(p_a) \approx 0.726$ and $i^\star(p_a) \approx 0.500$, giving ARE $\approx 68.9\%$.

Corollary 5. *The relative efficiency loss of the convex restriction is*

$$1 - \frac{i^\star(p_0)}{i(p_0)} = \frac{\mathrm{dist}_{L_2(P_0)}(\psi_0,\, \Psi_\downarrow(p_0))^2}{i(p_0)}.$$

In particular, for any non-increasing $g \in L_2(P_0)$,

$$1 - \frac{i^\star(p_0)}{i(p_0)} \leq \frac{\|\psi_0 - g\|^2_{L_2(P_0)}}{i(p_0)}.$$

This identity gives a geometric characterization of the efficiency gap: the convex restriction is costly precisely when the true score ψ_0 is far from the monotone cone in $L_2(P_0)$. "Near full efficiency" is equivalent to "the true score is close to being non-increasing."

Remark 3 (Numerical computations)**.** *The numerical values reported in the figures—$i(p_a) \approx 0.726$, $i^\star(p_a) \approx 0.500$, and* ARE $\approx 68.9\%$ *for the bimodal mixture with $a = 2$, as well as $i^\star(p_0) \approx 0.439$ and* ARE $\approx 87.8\%$ *for the Cauchy—were computed as follows. The Fisher information $i(p_a) = \int \psi_a(z)^2 p_a(z)\, dz$ was evaluated by adaptive numerical quadrature (Simpson's rule on $[-20, 20]$). The antitonic projection $\psi_0^\star$ was obtained by applying the Pool Adjacent Violators Algorithm (PAVA)*

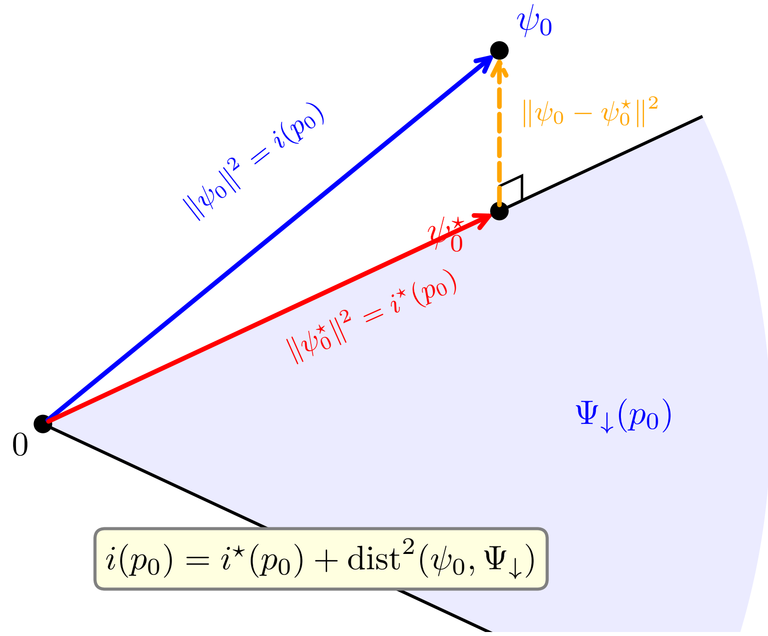

Figure 3: Geometric illustration of the Moreau decomposition in $L_2(P_0)$ Hilbert space. The closed convex cone $\Psi_\downarrow(p_0)$ (shaded wedge) contains all non-increasing score functions. The true score ψ_0 (blue arrow) lies outside the cone for non-log-concave p_0. Its $L_2(P_0)$-projection $\psi_0^\star$ (red arrow) onto the cone defines the antitonic information $i^\star(p_0) = \|\psi_0^\star\|^2$. The residual $\psi_0 - \psi_0^\star$ (orange dashed arrow) is orthogonal to $\psi_0^\star$, yielding the Pythagorean identity $i(p_0) = i^\star(p_0) + \mathrm{dist}^2(\psi_0, \Psi_\downarrow)$. When p_0 is log-concave, $\psi_0 \in \Psi_\downarrow(p_0)$ and the gap vanishes.

to the score ψ_a evaluated on a fine grid of 10,000 equally spaced points, weighted by the density p_a; the antitonic information $i^\star(p_a) = \int(\psi_0^\star)^2\, dP_0$ was then computed by numerical integration over the same grid. The Python code used for these computations, including the figure generation script, is available at `generate_figures.py` in the supplementary materials.

6 Open Problems

Three natural directions remain.

(i) **Attainability under minimal assumptions.** Our counterexample uses Yao–Zhao's KDRE, which requires finite third moments and smooth error densities. Proving that the full semiparametric bound is attainable under only finite Fisher information and Feng et al.'s conditions A1–A5–which allow heavy-tailed errors such as the Cauchy–remains open. The key missing step is a local translation expansion for the unprojected score estimator; see van der Vaart [1998] for the relevant Z-estimation framework.

(ii) **Explicit gap bounds for structured classes.** The identity (8) is abstract. Computing or bounding $\mathrm{dist}_{L_2(P_0)}(\psi_0, \Psi_\downarrow(p_0))^2$ for concrete non-log-concave families–Gaussian scale mixtures, Student's t_ν, contaminated normals, bounded-support densities–would give practical guidance

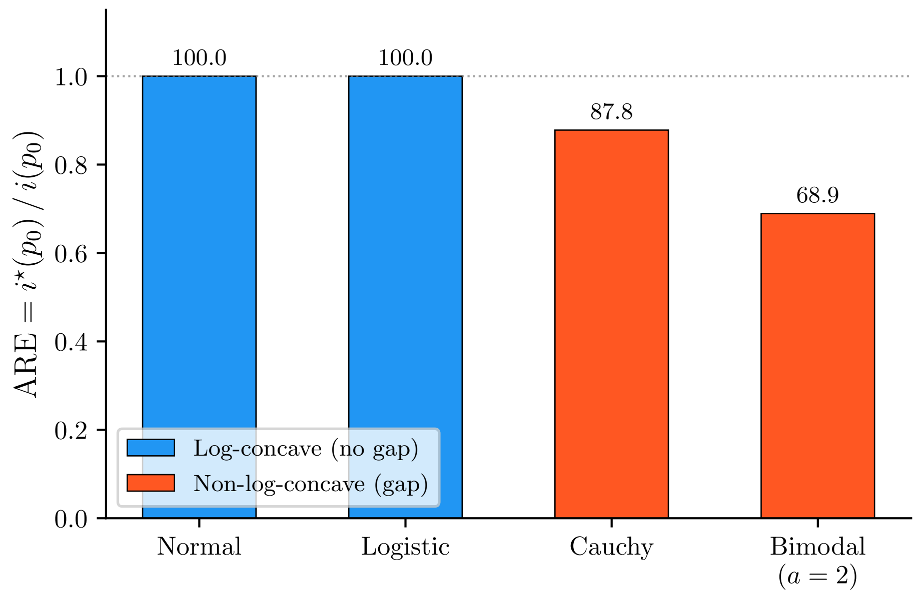

Figure 4: Asymptotic relative efficiency $\text{ARE} = i^\star(p_0)/i(p_0)$ of the score-matched optimal convex estimator relative to the full semiparametric bound, for four error densities. Log-concave densities (Normal, Logistic; blue) achieve $\text{ARE} = 100\%$: the convex restriction is costless. Non-log-concave densities (Cauchy, bimodal Gaussian mixture with $a = 2$; orange) have $\text{ARE} < 100\%$. For Cauchy errors, $\text{ARE} \approx 87.8\%$, reflecting a modest 12.2% efficiency loss. For the bimodal mixture, $\text{ARE} \approx 68.9\%$, a more substantial gap driven by the strongly non-monotone inter-modal region of the score function (cf. Figure 2).

on when the convex restriction is nearly costless. SolveAll lists this as a separate open problem [SolveAll.org, 2026a].

(iii) **Extensions beyond fixed dimension.** The present analysis is for fixed p. Growing-dimensional settings, heteroskedastic errors, dependent designs, and model misspecification are all natural extensions where the convex-versus-efficient comparison is open.

7 Conclusion

We have shown that the score-matched optimal convex M-estimator of Feng et al. [2026] attains the full semiparametric efficiency bound for the slope parameter in the linear regression model with unknown error density if and only if the error density is log-concave. For non-log-concave error densities, a strict efficiency gap exists, quantified exactly by the squared $L_2(P_0)$-distance from the true score to the monotone cone. A concrete counterexample with bimodal Gaussian mixture errors and Rademacher design demonstrates that the gap is real and that existing nonconvex adaptive estimators can exploit it.

It is worth noting that this asymptotic efficiency gap does not diminish the practical value of the convex estimator. Nonconvex alternatives such as the KDRE of Yao and Zhao [2013] require bandwidth selection, smooth error densities, and finite third moments, and may suffer from instability or local optima in finite samples. The convex restriction trades a quantifiable amount of asymptotic efficiency–zero for log-concave errors, and exactly $\text{dist}_{L_2(P_0)}(\psi_0, \Psi_\downarrow(p_0))^2/i(p_0)$ in relative terms–for guaranteed convexity of the optimization landscape, robustness to heavy tails (as illustrated by the Huber-like optimal loss for Cauchy errors), and the computational stability of the PAVA-based procedure.

References

Rudolf Beran. An efficient and robust adaptive estimator of location. *The Annals of Statistics*, 6(2): 292–313, 1978.

Peter J. Bickel. On adaptive estimation. *The Annals of Statistics*, 10(3):647–671, 1982.

Oliver Y. Feng, Yu-Chun Kao, Min Xu, and Richard J. Samworth. Optimal convex M-estimation via score matching. *The Annals of Statistics*, 54(1):408–441, 2026. arXiv:2403.16688v2.

Piet Groeneboom and Geurt Jongbloed. *Nonparametric Estimation under Shape Constraints*. Cambridge University Press, 2014.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2nd edition, 2009.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Stephen Portnoy and Roger Koenker. Adaptive L-estimation for linear models. *The Annals of Statistics*, 17(1):362–381, 1989.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

SolveAll.org. Quantify and characterize the efficiency gap induced by convex loss restriction, 2026a. Available at `https://solveall.org`.

SolveAll.org. Does the score-matched optimal convex estimator attain the full semiparametric efficiency bound?, 2026b. Available at `https://solveall.org`.

Charles Stein. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 187–195, 1956.

Charles J. Stone. Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, 3(2):267–284, 1975.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Constance van Eeden. Efficiency-robust estimation of location. *The Annals of Mathematical Statistics*, 41(1):172–181, 1970.

Weixin Yao and Zhibiao Zhao. Kernel density-based linear regression estimate. *Communications in Statistics—Theory and Methods*, 42(24):4499–4512, 2013. doi: 10.1080/03610926.2011.650269.