The Annals of Statistics
2026, Vol. 54, No. 1, 408–441
https://doi.org/10.1214/25-AOS2572

OPTIMAL CONVEX *M*-ESTIMATION VIA SCORE MATCHING

BY OLIVER Y. FENG[1,3,a], YU-CHUN KAO[2,b], MIN XU[2,c] AND RICHARD J. SAMWORTH[3,d]

[1]*Department of Mathematical Sciences, University of Bath,* [a]*of402@bath.ac.uk*
[2]*Department of Statistics, Rutgers University,* [b]*yk495@scarletmail.rutgers.edu,* [c]*mx76@stat.rutgers.edu*
[3]*Statistical Laboratory, University of Cambridge,* [d]*r.samworth@statslab.cam.ac.uk*

In the context of linear regression, we construct a data-driven convex loss function with respect to which empirical risk minimisation yields optimal asymptotic variance in the downstream estimation of the regression coefficients. At the population level, the negative derivative of the optimal convex loss is the best decreasing approximation of the derivative of the log-density of the noise distribution. This motivates a fitting process via a nonparametric extension of score matching, corresponding to a log-concave projection of the noise distribution with respect to the Fisher divergence. At the sample level, our semiparametric estimator is computationally efficient, and we prove that it attains the minimal asymptotic covariance among all convex M-estimators. As an example of a non-log-concave setting, the optimal convex loss function for Cauchy errors is Huber-like, and our procedure yields asymptotic efficiency greater than 0.87 relative to the maximum likelihood estimator of the regression coefficients that uses oracle knowledge of this error distribution. In this sense, we provide robustness and facilitate computation without sacrificing much statistical efficiency. Numerical experiments using our accompanying `R` package `asm` confirm the practical merits of our proposal.

1. Introduction. In linear models, the Gauss–Markov theorem is the primary justification for the use of ordinary least squares (OLS) in settings where the Gaussianity of our error distribution may be in doubt. It states that, provided the errors have a finite second moment, OLS attains the minimal covariance among all linear unbiased estimators; recent papers on this topic include Hansen (2022), Pötscher and Preinerstorfer (2024) and Lei and Wooldridge (2022). On the other hand, it is now understood that biased, nonlinear estimators can achieve lower mean squared error than OLS (Stein (1956a), Hoerl and Kennard (1970)), especially when the noise distribution is appreciably non-Gaussian (Zou and Yuan (2008), Dümbgen, Samworth and Schuhmacher (2011)). However, it remains unclear how best to fit linear models in a computationally efficient and adaptive fashion, that is, without knowledge of the error distribution.

Consider a linear model where $Y_i = X_i^\top \beta_0 + \varepsilon_i$ for $i = 1, \dots, n$. Recall that an *M-estimator* of $\beta_0 \in \mathbb{R}^d$ based on a loss function $\ell\colon \mathbb{R} \to \mathbb{R}$ is defined as an empirical risk minimiser

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i - X_i^\top \beta), \tag{1}$$

provided that this exists. If ℓ is differentiable on $\mathbb{R}$ with negative derivative $\psi = -\ell'$, then $\hat{\beta} \equiv \hat{\beta}_\psi$ solves the corresponding estimating equations

$$\frac{1}{n} \sum_{i=1}^{n} X_i \psi(Y_i - X_i^\top \hat{\beta}_\psi) = 0 \tag{2}$$

Received April 2024; revised May 2025.
MSC2020 subject classifications. Primary 62G35; secondary 62J05.
Key words and phrases. M-estimation, robust statistics, log-concavity, antitonic, score matching.

and is referred to as a *Z-estimator*. We study a random design setting in which the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed, with $X_1, \ldots, X_n$ being $\mathbb{R}^d$-valued covariates that are independent of real-valued errors $\varepsilon_1, \ldots, \varepsilon_n$ with density p_0. Suppose further that $\mathbb{E}\{X_1 \psi(\varepsilon_1)\} = 0$. This means that $\hat{\beta}_\psi$ is *Fisher consistent* in the sense that the population analogue of (2) is satisfied by the true parameter β_0, that is, $\mathbb{E}\{X_1 \psi(Y_1 - X_1^\top \beta_0)\} = 0$. Then under suitable regularity conditions, including ψ being differentiable and $\mathbb{E}(X_1 X_1^\top) \in \mathbb{R}^{d \times d}$ being invertible, we have

$$\begin{aligned} &\sqrt{n}(\hat{\beta}_\psi - \beta_0) \xrightarrow{d} N_d\big(0, V_{p_0}(\psi) \cdot \{\mathbb{E}(X_1 X_1^\top)\}^{-1}\big) \\ &\text{as } n \to \infty, \text{ where } V_{p_0}(\psi) := \frac{\mathbb{E}\psi^2(\varepsilon_1)}{\{\mathbb{E}\psi'(\varepsilon_1)\}^2} \end{aligned} \tag{3}$$

(e.g., van der Vaart (1998), Theorems 5.21, 5.23 and 5.41). Since the covariates and errors are assumed to be independent, they contribute separately to the limiting covariance above (a special case of the "sandwich" formula (Huber (1967), Young and Shah (2024))): the matrix $\{\mathbb{E}(X_1 X_1^\top)\}^{-1}$ depends only on the covariate distribution, whereas the scalar $V_{p_0}(\psi)$ depends on the loss function ℓ (through $\psi = -\ell'$) and on the error distribution.

If the errors $\varepsilon_1, \varepsilon_2, \ldots$ have a known absolutely continuous density p_0 on $\mathbb{R}$, then we can define the maximum likelihood estimator $\hat{\beta}^{\mathrm{MLE}}$ by taking $\ell = -\log p_0$ in (1). In this case, $\psi = -\ell'$ is the *score function (for location)*[1] $\psi_0 := (p_0'/p_0)\mathbb{1}_{\{p_0>0\}}$. Under appropriate regularity conditions (e.g., van der Vaart (1998), Theorem 5.39), including that the *Fisher information (for location)* $i(p_0) := \int_{\mathbb{R}} \psi_0^2 p_0 = \int_{\{p_0>0\}} (p_0')^2/p_0$ is finite, we have

$$\sqrt{n}(\hat{\beta}^{\mathrm{MLE}} - \beta_0) \xrightarrow{d} N_d\bigg(0, \frac{\{\mathbb{E}(X_1 X_1^\top)\}^{-1}}{i(p_0)}\bigg) \tag{4}$$

as $n \to \infty$. The limiting covariance matrix $\{\mathbb{E}(X_1 X_1^\top)\}^{-1}/i(p_0)$ constitutes the usual efficiency lower bound (van der Vaart (1998), Chapter 8). In fact, it can be seen directly that $1/i(p_0)$ is the smallest possible value of the *asymptotic variance factor* $V_{p_0}(\psi)$ in the limiting covariance of $\sqrt{n}(\hat{\beta}_\psi - \beta_0)$ in (3). Indeed, by the Cauchy–Schwarz inequality,

$$V_{p_0}(\psi) = \frac{\int_{\mathbb{R}} \psi^2 p_0}{(\int_{\mathbb{R}} \psi' p_0)^2} = \frac{\int_{\mathbb{R}} \psi^2 p_0}{(\int_{\mathbb{R}} \psi p_0')^2} \geq \frac{1}{\int_{\{p_0>0\}} (p_0')^2/p_0} = \frac{1}{i(p_0)} \in (0, \infty) \tag{5}$$

whenever the integration by parts in the second step is justified, and equality holds if and only if there exists $\lambda \neq 0$ such that $\psi(\varepsilon_1) = \lambda \psi_0(\varepsilon_1)$ almost surely. This leads to an equivalent variational definition of the Fisher information; see Huber and Ronchetti (2009), Theorem 4.2, which we restate as Proposition S25 in Section S4. Thus, when (4) holds, $\hat{\beta}^{\mathrm{MLE}}$ has minimal asymptotic covariance among all Z-estimators $\hat{\beta}_\psi$ for which (3) is valid, with the score function ψ_0 being the optimal choice of ψ.

Our goal in this work is to choose ψ in a data-driven manner, such that the corresponding loss function ℓ in (1) is convex, and such that the scale factor $V_{p_0}(\psi)$ in the asymptotic covariance (3) of the downstream estimator of β_0 is minimised. Convexity is a particularly convenient property for a loss function, since for the purpose of M-estimation, it leads to more tractable theory and computation. Indeed, the empirical risk in (1) becomes convex in β, so its local minimisers are global minimisers. In particular, when ℓ is also differentiable, $\hat{\beta}_\psi$ is a Z-estimator satisfying (2) if and only if it is an M-estimator satisfying (1).

[1] The score is usually defined as a function of a parameter $\theta \in \mathbb{R}$ as the derivative of the log-likelihood; the link with our terminology comes from considering the location model $\{p_0(\cdot + \theta) : \theta \in \mathbb{R}\}$, and evaluating the score at the origin.

The existence, uniqueness and $\sqrt{n}$-consistency of $\hat{\beta}_\psi$ are then guaranteed under milder conditions on ℓ than for generic loss functions (Yohai and Maronna (1979), Maronna and Yohai (1981), Portnoy (1985), Mammen (1989), Arcones (1998), He and Shao (2000)); see Proposition S24. Furthermore, an important practical advantage is that we can compute $\hat{\beta}_\psi$ efficiently using convex optimisation algorithms with guaranteed convergence (Boyd and Vandenberghe (2004), Chapter 9).

In view of the discussion above, our first main contribution in Section 2 is to determine the optimal population-level convex loss function in the sense described in the previous paragraph. For a uniformly continuous error density p_0, this amounts to finding

$$\psi_0^* \in \operatorname*{argmin}_{\psi \in \Psi_\downarrow(p_0)} V_{p_0}(\psi), \tag{6}$$

where $\Psi_\downarrow(p_0)$ denotes the class of decreasing, right-continuous functions ψ satisfying $\int_{\mathbb{R}} \psi^2 p_0 < \infty$. We will actually define the ratio $V_{p_0}(\psi)$ in a slightly more general way than in (5) to allow us to handle nondifferentiable functions ψ. This turns out to be convenient because, for instance, the robust Huber loss ℓ_K given by

$$\ell_K(z) := \begin{cases} z^2/2 & \text{if } |z| \le K, \\ K|z| - K^2/2 & \text{if } |z| > K, \end{cases} \tag{7}$$

for $K \in (0, \infty)$ has a nondifferentiable negative derivative $\psi_K := -\ell_K'$ satisfying $\psi_K(z) = (-K) \vee (-z) \wedge K$ for $z \in \mathbb{R}$.

In Section 2.1, we show that minimising $V_{p_0}(\cdot)$ over $\Psi_\downarrow(p_0)$ is equivalent to minimising the *score matching* objective

$$D_{p_0}(\psi) := \mathbb{E}\{\psi^2(\varepsilon_1) + 2\psi'(\varepsilon_1)\}, \tag{8}$$

over $\psi \in \Psi_\downarrow(p_0)$, provided that we take appropriate care in defining this expression when ψ is not absolutely continuous. This observation allows us to obtain an explicit characterisation of the solution to the optimisation problem as a "projected" score function ψ_0^* in terms of p_0 and its distribution function F_0. Indeed, to obtain ψ_0^* at $z \in \mathbb{R}$, we can first consider $p_0 \circ F_0^{-1}$ (whose domain is $[0, 1]$), then compute the right derivative of its least concave majorant, before finally applying the resulting function to $F_0(z)$. The negative antiderivative ℓ_0^* of ψ_0^* is then the optimal convex loss function we seek. An important property is that $\mathbb{E}\psi_0^*(Y_1 - X_1^\top \beta_0) = 0$, which ensures that ℓ_0^* correctly identifies the estimand β_0 on the population level; equivalently, $\hat{\beta}_{\psi_0^*}$ is Fisher consistent.

Note that $\hat{\beta}^{\mathrm{MLE}}$ is a convex M-estimator if and only if $\ell = -\log p_0$ is convex, that is, p_0 is log-concave, in which case $\psi_0^* = \psi_0$ by (5). We will be especially interested in error densities p_0 that are not log-concave, for which the efficiency lower bound in (5) cannot be achieved by a convex M-estimator corresponding to a decreasing function ψ. We interpret the minimum ratio $V_{p_0}(\psi_0^*)$ as an analogue of the inverse Fisher information, serving as the crucial part of the efficiency lower bound for convex M-estimators. To reinforce the link with score matching, we will see in Section 2.2 that the density proportional to $e^{-\ell_0^*}$ is the best log-concave approximation to p_0 with respect to the Fisher divergence defined formally in (22) below. This is typically different from the well-studied log-concave projection with respect to Kullback–Leibler divergence, and indeed the latter may yield considerably suboptimal covariance for the resulting convex M-estimator; see Proposition 7. In concrete examples where p_0 has heavy tails (e.g., a Cauchy density) or is multimodal (e.g., a mixture density), we compute closed-form expressions for the projected score function and the optimal convex loss function in Section 2.3. In particular, ℓ_0^* turns out to be a robust Huber-like loss function in the Cauchy case. Figure 1 presents plots of ψ_0^* and ℓ_0^* for three other distributions. More

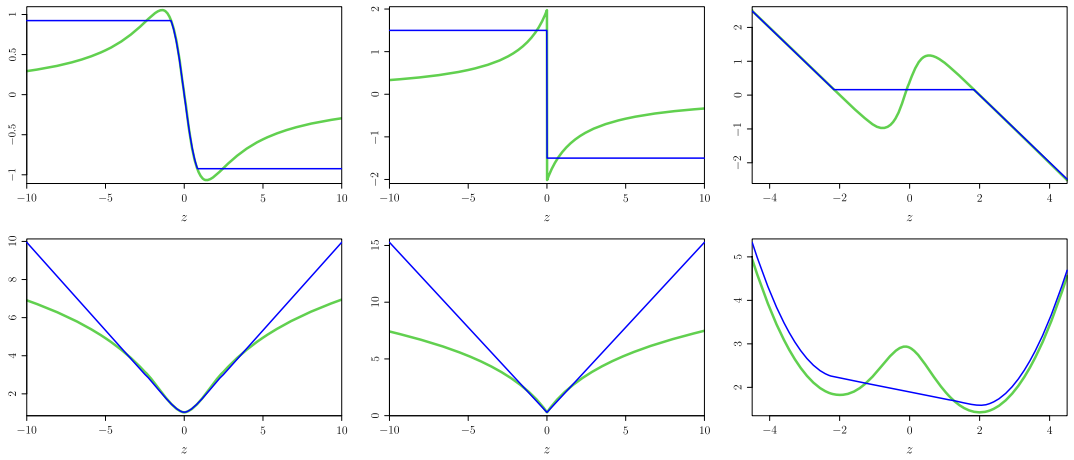

FIG. 1. *Top row*: *Plots of the score function* ψ_0 (*green*) *and projected score function* ψ_0^* (*blue*); *Bottom row*: *their respective negative antiderivatives*, *namely the negative log-density* $-\log p_0$ (*green*) *and optimal convex loss function* ℓ_0^* (*blue*), *for each of the following non-log-concave distributions (from left to right)*: *(a) Student's* t_2; *(b) symmetrised Pareto* (*S*25) *with* $\sigma = 2$ *and* $\alpha = 3$; *(c) Gaussian mixture* $0.4N(-2, 1) + 0.6N(2, 1)$.

generally, when the errors are heavy-tailed in the sense that their (two-sided) hazard function is bounded, Lemma 5 shows that the projected score function ψ_0^* is bounded, in which case the corresponding convex loss ℓ_0^* grows at most linearly in the tails, and hence is robust to outliers; see the discussion immediately preceding Lemma 5. A major advantage of our framework over the use of the Huber loss is that it does not require the choice of a transition point K (see (7) above) between quadratic and linear regimes (which in a regression context amounts to a choice of scale for the error distribution). In fact, our *antitonic*[2] score projection, and hence the Fisher divergence projection, is affine equivariant (Remark 9), which reflects the fact that we optimise $V_{p_0}(\cdot)$ in (6) over a class $\Psi_\downarrow(p_0)$ that is closed under multiplication by nonnegative scalars.

In Section 3, we turn our attention to a linear regression setting where the error density p_0 is unknown. We aim to construct a semiparametric M-estimator of β_0 that achieves minimal covariance among all convex M-estimators, but since the optimal loss function ℓ_0^* is unknown, we seek to estimate β_0 and ψ_0^* simultaneously. Our alternating optimisation procedure starts with a nonadaptive initialiser $\bar{\beta}_n$ and computes a kernel density estimate of the error distribution based on the residuals. We can then apply the linear-time Pool Adjacent Violators Algorithm (PAVA) to obtain the projected score function of the density estimate, before minimising its negative antiderivative using Newton-type optimisation techniques to yield an updated estimator. This process could then be iterated to convergence, but if we initialise with a $\sqrt{n}$-consistent pilot estimator $\bar{\beta}_n$, then one iteration of the alternating algorithm above suffices for our theoretical guarantees and, moreover, it ensures that the procedure is computationally efficient. When p_0 is symmetric, we prove that a three-fold cross-fitting version of our algorithm (with the different steps computed on different folds) yields an estimator $\hat{\beta}_n$ that is $\sqrt{n}$-consistent and asymptotically normal, with limiting covariance attaining our efficiency lower bound for convex M-estimators; see Theorem 13 in Section 3.3. We develop analogous methodology and theory for the setting where an explicit intercept is present in the linear model, and where the errors are appropriately centred though not necessarily symmetric; see Theorem 14 in Section 3.4. Consistent estimation of the limiting covariance

[2]Antitonic means decreasing, in contrast to isotonic (increasing) (Groeneboom and Jongbloed (2014), Section 2.1).

matrices is straightforward using our nonparametric score matching procedure, so combining this with our asymptotic distributional results for $\hat{\beta}_n$, we can then perform inference for β_0 (Section 3.5).

Section 4 is devoted to a numerical study of the empirical performance and computational efficiency of our antitonic score matching estimator, which is implemented in the `R` package `asm` (Kao et al. (2024)), and whose output is designed to mimic that of the standard existing `lm` function in several aspects so as to appear familiar to practitioners. These corroborate our theoretical findings: our proposed approach achieves (sometimes dramatically) smaller estimation error compared with alternatives such as OLS, the least absolute deviation (LAD) estimator, a semiparametric one-step estimator and a semiparametric M-estimator based on the log-concave MLE of the noise distribution. Moreover, the corresponding confidence sets for β_0 are smaller, while retaining nominal coverage. Finally, we perform a runtime analysis to show that the improved statistical performance comes without sacrificing computational scalability.

The proofs of the results in Sections 2 and 3 are deferred to Sections S1 and S3, respectively, of the Supplementary Material (Feng et al. (2026)). This also contains additional examples and background for Sections 2 and 3 together with further simulation results. All labels and headings in the supplement are prefixed with the letter "S".

1.1. *Related work.* Score matching (Hyvärinen (2005), Lyu (2012)) is an estimation method designed for statistical models where the likelihood is only known up to a normalisation constant (e.g., a partition function) that may be infeasible to compute; see the recent tutorial by Song and Kingma (2021) on "energy-based" models. Instead of maximising an approximation to the likelihood, score matching circumvents this issue altogether by estimating the derivative of a log-density, that is, the score function. More precisely, given a differentiable density p_0 on $\mathbb{R}^d$ with score function $\psi_0 := (\nabla p_0/p_0)\mathbb{1}_{\{p_0>0\}}$, the population version of the procedure aims to minimise

$$\mathbb{E}\big(\|\psi(\varepsilon) - \psi_0(\varepsilon)\|^2\big) \tag{9}$$

over a suitable class Ψ of differentiable functions $\psi \equiv (\psi_1, \ldots, \psi_d)\colon \mathbb{R}^d \to \mathbb{R}^d$, where $\varepsilon \sim p_0$. Hyvärinen (2005) used integration by parts to show that it is equivalent to minimise

$$D_{p_0}(\psi) := \mathbb{E}\big\{\|\psi(\varepsilon)\|^2 + 2(\nabla \cdot \psi)(\varepsilon)\big\} \tag{10}$$

over $\psi \in \Psi$, where $\nabla \cdot \psi := \sum_{j=1}^d \partial\psi_j/\partial x_j$. The score matching estimator based on data $\varepsilon_1, \ldots, \varepsilon_n$ in $\mathbb{R}^d$ is then defined as a minimiser of the empirical analogue $\hat{D}_n(\psi) := n^{-1}\sum_{i=1}^n\{\|\psi(\varepsilon_i)\|^2 + 2(\nabla \cdot \psi)(\varepsilon_i)\}$ over $\psi \in \Psi$; see also Cox (1985). Such estimators are important in the context of Langevin Monte Carlo (Parisi (1981), Roberts and Tweedie (1996), Betancourt et al. (2017), Cheng et al. (2018)) and diffusion models (Li et al. (2024)). The appearance of the score function in the (reverse-time) stochastic differential equations governing the Langevin and diffusion model dynamics can be related to Tweedie's formula, which underpins empirical Bayes denoising (Efron (2011), Derenski et al. (2023)).

Likelihood maximisation corresponds to distributional approximation with respect to the Kullback–Leibler divergence. On the other hand, score matching seeks to minimise the *Fisher divergence* (Johnson (2004), Section 1.3) from a class of densities to the target p_0, in view of the equivalence between the optimisation objectives (9) and (10); see (20) below. Sriperumbudur et al. (2017) studied infinite-dimensional exponential families indexed by reproducing kernel Hilbert spaces, and proposed and analysed a density estimator that minimises a penalised empirical Fisher divergence. Koehler, Heckett and Risteski (2022) used isoperimetric inequalities to investigate the statistical efficiency of score matching relative to maximum

likelihood, thereby quantifying the effect of eliminating normalisation factors. Lyu (2012) observed that Fisher divergence and Kullback–Leibler divergence are related by an analogue of de Bruijn's identity (Johnson (2004), Appendix C; Cover and Thomas (2006), Section 17.7), which links Fisher information and Shannon entropy. From an information-theoretic perspective, Johnson and Barron (2004) proved central limit theorems that establish convergence in Fisher divergence to a limiting Gaussian distribution. Ley and Swan (2013) extended Stein's method to derive information inequalities that bound a variety of integral probability distances in terms of the Fisher divergence.

Score matching has been generalised in different directions and applied to a variety of statistical problems (e.g., Hyvärinen (2007), Vincent (2011), Lyu (2012), Mardia, Kent and Laha (2016), Song et al. (2020), Yu, Gupta and Kolar (2020), Yu, Drton and Shojaie (2022), Lederer and Oesting (2023), Benton et al. (2024), Ghosh et al. (2025)), where it exhibits excellent empirical performance while being computationally superior to full likelihood approaches. In particular, score-based algorithms for generative modelling, via Langevin dynamics (Song and Ermon (2019)) and diffusion models (Song et al. (2021)), have achieved remarkable success in machine learning tasks such as the reconstruction, inpainting and artificial generation of images; see, for example, Jolicoeur-Martineau et al. (2020), De Bortoli et al. (2022) and many other references therein. In these applications, score matching is applied to a class of functions parametrised by the weights of a deep neural network. On the other hand, different statistical considerations lead us to develop a nonparametric extension of score matching in Section 2, which we use to construct data-driven convex loss functions for efficient semiparametric estimation. We see that it is by minimising the Fisher divergence instead of the Kullback–Leibler divergence to the error distribution that one obtains a convex M-estimator with minimal asymptotic variance.

The framework in Section 3.3 includes as a special case the classical location model in which we observe $Y_i = \theta_0 + \varepsilon_i$ for $i = 1, \dots, n$, where $\theta_0 \in \mathbb{R}$ is the parameter of interest and $\varepsilon_1, \dots, \varepsilon_n$ are independent errors with an unknown density p_0 that is symmetric about 0. Starting from the seminal paper of Stein (1956b), a series of works (e.g., van Eeden (1970), Stone (1975), Beran (1978), Bickel (1982), Schick (1986), Faraway (1992), Dalalyan, Golubev and Tsybakov (2006), van der Vaart and Wellner (2021), Gupta, Lee and Price (2023)) showed that adaptive, asymptotically efficient estimators of θ_0 can be constructed; see also Doss and Wellner (2019) and Laha (2021) for approaches based on the further assumption that p_0 is log-concave. Many of these traditional semiparametric procedures have drawbacks that limit their practical utility. In particular, the estimated likelihood may have multiple local optima and it may be difficult to guarantee convergence of an optimisation algorithm to a global maximum (van der Vaart (1998), Example 5.50). This is one of the reasons why prior works often study a one-step estimator resulting from a single iteration of Newton's method (Bickel (1975), Jin (1990), Mammen and Park (1997), Laha (2021)), rather than full likelihood maximisation, though finite-sample performance may remain poor and sensitive to tuning (see Section 4). By contrast, our focus is not on classical semiparametric adaptive efficiency per se; instead, we directly study the theoretical properties of a minimiser of the empirical risk with respect to an estimated loss function, whose convexity ensures that the estimator can be computed efficiently by iterating gradient descent or Newton's method to convergence.

Recently, Kao, Xu and Zhang (2024) constructed a location M-estimator that can adaptively attain rates of convergence faster than $n^{-1/2}$ when the symmetric error density is compactly supported and suitably irregular (e.g., discontinuous at the boundary of its support). They considered ℓ^q-location estimators $\hat{\theta}_q := \operatorname{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^n |Y_i - \theta|^q$ based on univariate observations $Y_1, \dots, Y_n$, and used Lepski's method to select an exponent $\hat{q} \in [2, \infty)$ that minimises a proxy for the asymptotic variance of $\hat{\theta}_q$. The resulting estimator $\hat{\theta}_{\hat{q}}$ is shown to

be minimax optimal up to polylogarithmic factors, and the procedure is extended to linear regression models with unknown symmetric errors. By comparison with Kao, Xu and Zhang (2024), we study "regular" regression models where the Fisher information is finite and minimax rates faster than $n^{-1/2}$ are impossible to achieve. We aim to minimise the asymptotic variance as an end in itself, over the entire nonparametric class of convex loss functions rather than ℓ^q-loss functions specifically.

Our work has connections with robust statistics, which deals with heavy-tailed noise distributions and data that may be contaminated by random or adversarial outliers. As mentioned previously, robust loss functions are designed to be tolerant to such data corruption; examples include the Huber loss (7), a two-parameter family of loss functions considered by Barron (2019) and the antiderivatives of Catoni's influence functions (Catoni (2012)). The Huber loss functions ℓ_K originally arose because they have the minimax asymptotic variance property that for every $\epsilon \in (0,1)$, there exists a unique $K \equiv K_\epsilon > 0$ such that

$$\psi_K \equiv -\ell_K' = \operatorname*{argmin}_{\psi\in\Psi} \sup_{P\in\mathcal{P}_\epsilon^{\mathrm{sym}}(\Phi)} \underbrace{\frac{\int_{\mathbb{R}} \psi^2\,dP}{\int_{\mathbb{R}} \psi'\,dP}}_{=:V_P(\psi)},$$

where Ψ consists of all "sufficiently regular" $\psi : \mathbb{R} \to \mathbb{R}$, and the symmetric *$\epsilon$-contamination neighbourhood* $\mathcal{P}_\epsilon^{\mathrm{sym}}(\Phi)$ contains all univariate probability distributions of the form $P = (1-\epsilon)N(0,1) + \epsilon Q$ for some symmetric distribution Q. The pioneering paper of Huber (1964) also developed variational theory for minimising $\sup_{P\in\mathcal{P}} V_P(\cdot)$ more generally when $\mathcal{P}$ is a convex class of distributions, such as an ϵ-contamination or a Kolmogorov neighbourhood of a symmetric log-concave density (Huber and Ronchetti (2009), Section 4.5). See Donoho and Montanari (2015) for a high-dimensional extension of this line of work. An alternative to the Huber loss that seeks robustness without serious efficiency loss relative to OLS is the composite quantile regression (CQR) estimator (Zou and Yuan (2008), Yang and Wang (2024)); in fact, our approach is always at least as efficient as CQR (see Lemma S23). Other recent papers on robust convex M-estimation include Chinot, Lecué and Lerasle (2020) and Brunel (2023); see also the notes on robust statistical learning theory by Lerasle (2019).

More closely related to our optimisation problem (6) is the work of Hampel (1974) on *optimal B-robust estimators*, which have minimal asymptotic variance subject to an upper bound on the *gross error sensitivity* (Hampel et al. (1986), Section 2.4). In our linear regression setting with $\varepsilon_1 \sim p_0$, this amounts to

$$\text{minimising } V_{p_0}(\psi) \text{ over all "regular" } \psi \text{ such that } \int_{\mathbb{R}} \psi p_0 = 0 \text{ and } \sup_{z\in\mathbb{R}} |\psi(z)| \le b \tag{11}$$

for some suitable $b > 0$ (Hampel et al. (1986), p. 121 and Section 2.5d). In particular, when p_0 is a standard Gaussian density, ψ_K is again optimal for some $K \equiv K_b > 0$ that depends nonlinearly on b. By contrast with (6) however, the Fisher consistency condition $\mathbb{E}\psi(\varepsilon_1) = 0$ must be explicitly included as a constraint in (11) and, moreover, the L^∞ bound on ψ means that the set of feasible ψ is not closed under nonnegative scalar multiplication. Consequently, the resulting optimal location M-estimators are generally not scale invariant (Hampel et al. (1986), p. 105). In robust regression, adaptive selection of scale parameters is a nontrivial problem (e.g., van der Vaart (1998), Section 5.4; Huber and Ronchetti (2009), Section 7.7; Loh (2021)); see also Figure 5 below.

Finally, we mention a different line of work on the performance of linear regression M-estimators in a proportional asymptotic regime where $n/d \to \kappa \in (1,\infty)$ and the covariates are Gaussian. Bean et al. (2013) derived the unpenalised convex M-estimator with minimal expected out-of-sample prediction error when the errors are log-concave. Here, the optimisation objective is no longer $V_{p_0}(\psi)$ but instead the solution to a pair of nonlinear equations

involving the proximal operator of the convex loss function (El Karoui et al. (2013), El Karoui (2018)). For general error distributions with finite variance, Donoho and Montanari (2016) established exact asymptotics for convex M-estimators by means of an approximate message passing algorithm. Under prior structural information about the entries of β_0, Celentano and Montanari (2022) obtained precise characterisations of the asymptotic ℓ_2-estimation error of convex-regularised least squared estimators, Bayes-optimal approximate message passing and the Bayes risk, quantifying the gaps between computational feasible and statistically optimal estimators.

1.2. *Notation.* Throughout this paper, we will adopt the convention $0/0 := 0$ and write $[n] := \{1, \dots, n\}$ for $n \in \mathbb{N}$. For a function $f : \mathbb{R} \to \mathbb{R}$, let $\|f\|_\infty := \sup_{z\in\mathbb{R}} |f(z)|$. Recall that f is *symmetric* (i.e., *even*) if $f(z) = f(-z)$ for all $z \in \mathbb{R}$, and *antisymmetric* (i.e., *odd*) if $f(z) = -f(-z)$ for all $z \in \mathbb{R}$. For an open set $U \subseteq \mathbb{R}$, we say that $f : U \to \mathbb{R}$ is *locally absolutely continuous* on U if it is absolutely continuous on every compact interval $I \subseteq U$; or equivalently, if there exists a measurable function $g : U \to \mathbb{R}$ such that for every compact subinterval $I \subseteq U$, we have $\int_I |g| < \infty$ and $f(z_2) = f(z_1) + \int_{z_1}^{z_2} g$ for all $z_1, z_2 \in I$. In this case, f is differentiable Lebesgue almost everywhere on U, with $f' = g$ almost everywhere.

Given a Borel probability measure P on $\mathbb{R}$, we write $L^2(P)$ for the set of all Lebesgue measurable functions f on $\mathbb{R}$ for which $\|f\|_{L^2(P)} := (\int_{\mathbb{R}} f^2\, dP)^{1/2} < \infty$. Denote by $\langle f, g\rangle_{L^2(P)} := \int_{\mathbb{R}} fg\, dP$ the $L^2(P)$-inner product of $f, g \in L^2(P)$.

For a function $F : [0, 1] \to \mathbb{R}$, we write $\hat{F}$ for its least concave majorant on $[0, 1]$. Denote by $F^{(\mathrm{L})}(u)$ and $F^{(\mathrm{R})}(u)$, respectively, the left and right derivatives of F at $u \in [0, 1]$, whenever these are well-defined. Given an integrable function $f : (0, 1) \to \mathbb{R}$ with antiderivative $F : [0, 1] \to \mathbb{R}$ given by $F(u) := \int_0^u f$, define $\widehat{\mathcal{M}}_{\mathrm{R}} f : [0, 1] \to [-\infty, \infty]$ by

$$(\widehat{\mathcal{M}}_{\mathrm{R}} f)(u) := \begin{cases} \hat{F}^{(\mathrm{R})}(u) & \text{for } u \in [0, 1), \\ \hat{F}^{(\mathrm{L})}(1) & \text{for } u = 1, \end{cases}$$

so that $(\widehat{\mathcal{M}}_{\mathrm{R}} f)(1) = \lim_{u\nearrow 1}(\widehat{\mathcal{M}}_{\mathrm{R}} f)(u)$ by Rockafellar (1997), Theorem 24.1. Furthermore, define $\widehat{\mathcal{M}}_{\mathrm{L}} f : [0, 1] \to [-\infty, \infty]$ by $(\widehat{\mathcal{M}}_{\mathrm{L}} f)(u) := (\widehat{\mathcal{M}}_{\mathrm{R}} g)(1 - u)$ for $u \in [0, 1]$, where $g(u) := f(1 - u)$ for all such u.

2. The antitonic score projection.

2.1. *Construction and basic properties.* The aim of this section is to define formally and solve the optimisation problem (6) that yields the minimal asymptotic covariance of a convex M-estimator of β_0. Let P_0 be a probability measure on $\mathbb{R}$ with a uniformly continuous density p_0, which necessarily satisfies $p_0(\pm\infty) := \lim_{z\to\pm\infty} p_0(z) = 0$. Letting $\operatorname{supp} p_0 := \{z \in \mathbb{R} : p_0(z) > 0\}$, define $\mathcal{S}_0 \equiv \mathcal{S}(p_0) := (\inf(\operatorname{supp} p_0), \sup(\operatorname{supp} p_0))$, which is the smallest open interval that contains $\operatorname{supp} p_0$. We write $\Psi_\downarrow(p_0)$ for the set of all $\psi \in L^2(P_0)$ that are decreasing and right continuous. Observe that $\Psi_\downarrow(p_0)$ is a convex cone, that is, $c_1\psi_1 + c_2\psi_2 \in \Psi_\downarrow(p_0)$ whenever $\psi_1, \psi_2 \in \Psi_\downarrow(p_0)$ and $c_1, c_2 \geq 0$. Moreover, every $\psi \in \Psi_\downarrow(p_0)$ is necessarily finite-valued on $\mathcal{S}_0$, so the corresponding Lebesgue–Stieltjes integral $\int_{\mathcal{S}_0} p_0\, d\psi \in [-\infty, 0]$ is well-defined.

For $\psi \in \Psi_\downarrow(p_0)$ with $\int_{\mathbb{R}} \psi^2\, dP_0 > 0$, let

$$V_{p_0}(\psi) := \frac{\int_{\mathbb{R}} \psi^2\, dP_0}{(\int_{\mathcal{S}_0} p_0\, d\psi)^2} \in [0, \infty], \tag{12}$$

where we have modified the denominator in (5) to extend the original definition to nondifferentiable functions in $\Psi_\downarrow(p_0)$ such as $z \mapsto -\operatorname{sgn}(z)$. That $V_{p_0}(\psi)$ is indeed the asymptotic variance factor for the corresponding convex M-estimator is justified formally by

Proposition S24. As a first step towards minimising $V_{p_0}(\psi)$ over $\psi \in \Psi_\downarrow(p_0)$, note that $V_{p_0}(c\psi) = V_{p_0}(\psi)$ for every $c > 0$, so any minimiser is at best unique up to a positive scalar. Ignoring unimportant edge cases where the denominator in (12) is zero or infinity, our optimisation problem can therefore be formulated as a constrained minimisation of the numerator in (12) subject to the denominator being equal to 1. This motivates the definition of

$$D_{p_0}(\psi) := \int_{\mathbb{R}} \psi^2 \, dP_0 + 2 \int_{\mathcal{S}_0} p_0 \, d\psi \in [-\infty, \infty) \tag{13}$$

for $\psi \in \Psi_\downarrow(p_0)$, which resembles a Lagrangian, though analogously to, for example, Silverman (1982), page 798 and Dümbgen, Samworth and Schuhmacher (2011), page 705, there is no need to introduce a Lagrange multiplier. If ψ is locally absolutely continuous on $\mathcal{S}_0$ with derivative ψ' Lebesgue almost everywhere, then

$$D_{p_0}(\psi) = \int_{\mathbb{R}} \psi^2 \, dP_0 + 2 \int_{\mathcal{S}_0} \psi' p_0 = \int_{\mathbb{R}} \bigl(\psi^2 + 2\psi'\bigr) \, dP_0 = \mathbb{E}\bigl(\psi^2(\varepsilon_1) + 2\psi'(\varepsilon_1)\bigr) \tag{14}$$

when $\varepsilon_1 \sim P_0$, which we recognise as the score matching objective (8) in the introduction.

The formal link between $V_{p_0}(\cdot)$ and $D_{p_0}(\cdot)$ is that for $\psi \in \Psi_\downarrow(p_0)$ with $\int_{\mathbb{R}} \psi^2 \, dP_0 > 0$, we have $\int_{\mathcal{S}_0} p_0 \, d\psi \le 0$ and $c\psi \in \Psi_\downarrow(p_0)$ for all $c \ge 0$, so

$$\inf_{c \ge 0} D_{p_0}(c\psi) = \inf_{c \ge 0} \biggl(c^2 \int_{\mathbb{R}} \psi^2 \, dP_0 + 2c \int_{\mathcal{S}_0} p_0 \, d\psi \biggr) = -\frac{(\int_{\mathcal{S}_0} p_0 \, d\psi)^2}{\int_{\mathbb{R}} \psi^2 \, dP_0} = -\frac{1}{V_{p_0}(\psi)}. \tag{15}$$

Thus, minimising $V_{p_0}(\cdot)$ over $\Psi_\downarrow(p_0)$ is equivalent to minimising $D_{p_0}(\cdot)$ up to a scalar multiple, but $D_{p_0}(\cdot)$ is a convex function that is more tractable than $V_{p_0}(\cdot)$.

By exploiting this connection with score matching together with ideas from monotone function estimation, we prove in Theorem 2 below that the solution to our asymptotic variance minimisation problem is the function ψ_0^* that we construct explicitly in the following lemma.

LEMMA 1. *Let P_0 be a distribution with a uniformly continuous density p_0 on $\mathbb{R}$. Let $F_0 \colon [-\infty, \infty] \to [0, 1]$ be the corresponding distribution function, and for $u \in [0, 1]$, define*

$$F_0^{-1}(u) := \inf\bigl\{z \in [-\infty, \infty] : F_0(z) \ge u\bigr\} \quad \text{and} \quad J_0(u) := \bigl(p_0 \circ F_0^{-1}\bigr)(u).$$

Then both J_0 and its least concave majorant $\hat{J}_0$ on $[0, 1]$ are continuous, with $p_0 = J_0 \circ F_0$ on $\mathbb{R}$. Moreover,

$$\psi_0^* := \hat{J}_0^{(\mathrm{R})} \circ F_0$$

is decreasing and right continuous as a function from $\mathbb{R}$ to $[-\infty, \infty]$, provided that we set $\hat{J}_0^{(\mathrm{R})}(1) := \lim_{u \nearrow 1} \hat{J}_0^{(\mathrm{R})}(u)$. We have $\psi_0^*(z) \in \mathbb{R}$ if and only if $z \in \mathcal{S}_0$.

We refer to J_0 as the *density quantile function* (Parzen (1979), Jones (1992)). In the case where p_0 is a standard Cauchy density, Figure 2 presents a visualisation of J_0 and its least concave majorant $\hat{J}_0$, as well as the corresponding score functions $\psi_0 = p_0'/p_0$ and ψ_0^*.

THEOREM 2. *In the setting of Lemma 1, the following statements hold*:

(a) $\int_{\mathbb{R}} \psi_0^* \, dP_0 = 0$.

(b) *Let $i^*(p_0) := \int_{\mathbb{R}} (\psi_0^*)^2 \, dP_0$. Then $\inf_{\psi \in \Psi_\downarrow(p_0)} D_{p_0}(\psi) = -i^*(p_0)$.*

(c) *Suppose that $i^*(p_0) < \infty$. Then ψ_0^* is the unique minimiser of $D_{p_0}(\cdot)$ over $\Psi_\downarrow(p_0)$.*

Moreover, for every $\psi \in \Psi_\downarrow(p_0)$ such that $\int_{\mathbb{R}} \psi^2 \, dP_0 > 0$, we have

$$V_{p_0}(\psi) \ge V_{p_0}\bigl(\psi_0^*\bigr) = \frac{1}{i^*(p_0)} \in (0, \infty), \tag{16}$$

with equality if and only if $\psi = \lambda \psi_0^*$ for some $\lambda > 0$.

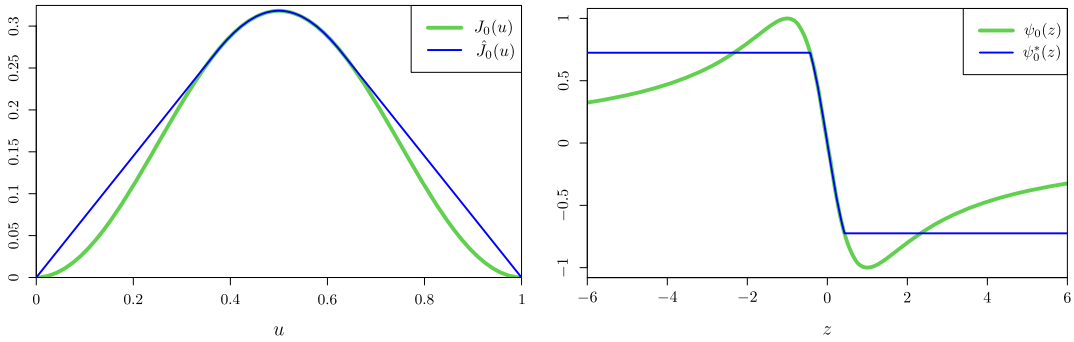

FIG. 2. Left: *The density quantile function J_0 and its least concave majorant $\hat{J}_0$ for a standard Cauchy density.* Right: *The corresponding score functions ψ_0 and ψ_0^*.*

(d) *Assume further that p_0 is absolutely continuous on $\mathbb{R}$ with derivative p_0' Lebesgue almost everywhere, corresponding score function*[3] $\psi_0 := p_0'/p_0$ *and Fisher information* $i(p_0) := \int_{\mathbb{R}} \psi_0^2 p_0$. *Then*

$$\psi_0^* = \widehat{\mathcal{M}}_{\mathrm{R}}(\psi_0 \circ F_0^{-1}) \circ F_0 \tag{17}$$

and $0 < i^*(p_0) \le i(p_0)$, *with equality if and only if p_0 is log-concave. In particular, if $i(p_0) < \infty$, then the conclusions of (c) hold.*

Some remarks on the proof and implications of this result are in order. The least concave majorant construction of ψ_0^* in Lemma 1 is reminiscent of isotonic regression, where the least squares estimator is the projection $\widehat{\theta} = \operatorname{argmin}_{\theta \in \mathcal{M}} \sum_{i=1}^n (Y_i - \theta_i)^2$ of the response vector $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ onto the monotone cone $\mathcal{M} := \{\theta = (\theta_1, \dots, \theta_n) : \theta_1 \le \cdots \le \theta_n\} \subseteq \mathbb{R}^n$. It is well known that the entries of $\widehat{\theta}$ are left derivatives of the greatest convex minorant of a cumulative sum diagram, and other monotonicity-constrained estimators (such as the Grenander estimator of a decreasing density) have similar explicit representations (e.g., Robertson, Wright and Dykstra (1988), Chapter 1; Groeneboom and Jongbloed (2014), Chapter 2; Samworth and Shah (2025), Chapter 9). An equivalent characterisation of the projection $\widehat{\theta}$ of Y onto the convex cone $\mathcal{M}$ is that

$$(Y - \widehat{\theta})^\top \theta \le 0 \quad \text{for all } \theta \in \mathcal{M} \quad \text{and} \quad (Y - \widehat{\theta})^\top \widehat{\theta} = 0. \tag{18}$$

To minimise the score matching objective $D_{p_0}(\cdot)$ over our convex cone $\Psi_\downarrow(p_0)$ of antitonic functions, the key step of the proof of Theorem 2 is to establish a similar first-order condition on the population level, namely

$$-\int_{\mathcal{S}_0} p_0 \, d\psi \le \int_{\mathbb{R}} \psi_0^* \psi \, dP_0 = \langle \psi_0^*, \psi \rangle_{L^2(P_0)} \tag{19}$$

for $\psi \in \Psi_\downarrow(p_0)$, with equality when $\psi = \psi_0^*$. Since both sides of (19) are linear in ψ, it suffices to prove it for indicator functions of the form $\psi = \mathbb{1}_{(-\infty,t]}$ for $t \in \mathbb{R}$, which generate the cone $\Psi_\downarrow(p_0)$; see (S5). This relies on key properties of the least concave majorant. Taking $\psi \equiv 1$ and $\psi \equiv -1$ in (19) yields $\mathbb{E}\psi_0^*(\varepsilon_1) = \langle \psi_0^*, 1 \rangle_{L^2(P_0)} = 0$. This is part *(a)* of the theorem, and ensures the Fisher consistency of the regression Z-estimator $\hat{\beta}_{\psi_0^*}$ defined in (2). The optimality properties of ψ_0^* in parts *(b)* and *(c)* follow readily from (19). In particular, combining this with the Cauchy–Schwarz inequality shows that ψ_0^* minimises the asymptotic

[3]Our convention $0/0 = 0$ means that $\psi_0 = (p_0'/p_0)\mathbb{1}_{\{p_0 > 0\}}$.

variance factor over $\Psi_\downarrow(p_0)$, similar to (5) in the introduction for the usual inverse Fisher information lower bound.

The parallels between (18) and (19) can be seen when p_0 is absolutely continuous with score function ψ_0 satisfying $i(p_0) = \|\psi_0\|^2_{L^2(P_0)} < \infty$. In this case, integration by parts yields $-\int_{S_0} p_0 \, d\psi = \int_{\mathbb{R}} \psi \psi_0 p_0$, and hence (19) states that $\langle \psi_0 - \psi_0^*, \psi \rangle_{L^2(P_0)} \le 0$ for $\psi \in \Psi_\downarrow(p_0)$ and $\langle \psi_0 - \psi_0^*, \psi_0^* \rangle_{L^2(P_0)} = 0$. Consequently,

$$D_{p_0}(\psi) = \int_{\mathbb{R}} (\psi - \psi_0)^2 \, dP_0 - \int_{\mathbb{R}} \psi_0^2 \, dP_0 = \|\psi - \psi_0\|^2_{L^2(P_0)} - i(p_0)$$

for all $\psi \in \Psi_\downarrow(p_0)$, so if $i(p_0) < \infty$, then

$$\psi_0^* \in \operatorname*{argmin}_{\psi \in \Psi_\downarrow(p_0)} D_{p_0}(\psi) = \operatorname*{argmin}_{\psi \in \Psi_\downarrow(p_0)} \|\psi - \psi_0\|^2_{L^2(P_0)}. \tag{20}$$

Thus, in the terminology of Section S5, ψ_0^* is a version of the $L^2(P_0)$-*antitonic projection* of ψ_0 onto $\Psi_\downarrow(p_0)$. Indeed, the explicit representation (17) of ψ_0^* as a "monotonisation" of ψ_0 (see the right panel of Figure 2) is consistent with that given in Proposition S33 for a general $L^2(P)$-antitonic projection, where P is a univariate probability measure with a continuous distribution function.

The inequality $i^*(p_0) = \int_{\mathbb{R}} (\psi_0^*)^2 \, dP_0 \le \int_{\mathbb{R}} \psi_0^2 \, dP_0 = i(p_0)$ in Theorem 2*(d)* follows from the fact that the $L^2(P_0)$-antitonic projection onto the convex cone $\Psi_\downarrow(p_0)$ is 1-Lipschitz with respect to $\|\cdot\|_{L^2(P_0)}$; see (S102) in Lemma S34. A statistical explanation of this information inequality arises from the fact that $1/i(p_0)$ is the infimum of the asymptotic variance factor $V_{p_0}(\psi)$ in (5) over all sufficiently regular $\psi : \mathbb{R} \to \mathbb{R}$; see Proposition S25. On the other hand, by (16), $1/i^*(p_0)$ is the minimum value of $V_{p_0}(\cdot)$ over the restricted class $\Psi_\downarrow(p_0)$, so in view of our discussion in the introduction, it can be interpreted as an information lower bound for convex M-estimators. When $i^*(p_0) < \infty$, the *antitonic relative efficiency*

$$\mathrm{ARE}^*(p_0) := \frac{i^*(p_0)}{i(p_0)}$$

therefore quantifies the price we pay in statistical efficiency for insisting that our loss function be convex. By Theorem 2*(d)*, $\mathrm{ARE}^*(p_0) \le 1$ with equality if and only if p_0 is log-concave, so we can regard $1 - \mathrm{ARE}^*(p_0)$ as a measure of departure from log-concavity; see Section 2.2 below. Example 12 shows that $\mathrm{ARE}^*(p_0) \approx 0.878$ when p_0 is the Cauchy density, whereas Example S7 in Section S2 yields a density p_0 for which $\mathrm{ARE}^*(p_0) = 0$. More generally, in Lemma 8 below, we provide a simple lower bound on $\mathrm{ARE}^*(p_0)$ that is reasonably tight for heavy-tailed densities p_0.

When p_0 is only uniformly continuous but not absolutely continuous, the score function and Fisher information cannot be defined as above, but we nevertheless refer to ψ_0^* and $i^*(p_0)$ as the *antitonic projected score function* (see Lemma 6 below) and *antitonic information (for location)*, respectively.

REMARK 3. Since F_0 is continuous, we have $(F_0 \circ F_0^{-1})(u) = u$ for all $u \in (0, 1)$. The concave function $\hat{J}_0$ is therefore absolutely continuous on $[0, 1]$ with derivative $\hat{J}_0^{(\mathrm{R})} = \psi_0^* \circ F_0^{-1}$ Lebesgue almost everywhere (Rockafellar (1997), Corollary 24.2.1), so

$$i^*(p_0) = \int_{\mathbb{R}} (\psi_0^*)^2 \, dP_0 = \int_0^1 (\psi_0^* \circ F_0^{-1})^2 = \int_0^1 (\hat{J}_0^{(\mathrm{R})})^2.$$

When p_0 is absolutely continuous on $\mathbb{R}$, a straightforward calculation (e.g., (S11) in the proof of Theorem 2) shows that the density quantile function $J_0 = p_0 \circ F_0^{-1}$ is absolutely

continuous on $[0, 1]$ with derivative $J_0' = \psi_0 \circ F_0^{-1}$ Lebesgue almost everywhere. Therefore, $\psi_0^* \circ F_0^{-1} = \hat{J}_0^{(\mathrm{R})} = \widehat{\mathcal{M}}_{\mathrm{R}}(\psi_0 \circ F_0^{-1})$ almost everywhere and

$$i(p_0) = \int_{\mathbb{R}} \psi_0^2 \, dP_0 = \int_0^1 (\psi_0 \circ F_0^{-1})^2 = \int_0^1 (J_0')^2.$$

It is well known that the map $p_0 \mapsto i(p_0)$ is convex on the space of absolutely continuous densities on $\mathbb{R}$ (e.g., Huber and Ronchetti (2009), Section 4.4, p. 78). The following corollary of Theorem 2*(b)* establishes that the analogous property holds for antitonic information.

COROLLARY 4. *The map $p_0 \mapsto i^*(p_0)$ is convex on the space of uniformly continuous densities on $\mathbb{R}$.*

Finally, in this subsection, we define the *two-sided hazard function* $h_0 \colon \mathbb{R} \to [0, \infty)$ of p_0 by

$$h_0(z) := \frac{p_0(z)}{F_0(z) \wedge (1 - F_0(z))} = \begin{cases} p_0(z)/F_0(z) & \text{if } F_0(z) \le 1/2, \\ p_0(z)/(1 - F_0(z)) & \text{if } F_0(z) > 1/2, \end{cases} \tag{21}$$

where in accordance with our convention $0/0 = 0$, we have $h_0(z) = 0$ whenever $F_0(z) \in \{0, 1\}$. The following simple lemma provides a necessary and sufficient condition on the two-sided hazard function for ψ_0^* to be appropriately bounded, which means that any negative antiderivative ℓ_0^* grows at most linearly in the tails. This is relevant because ψ_0^* is bounded if and only if the corresponding M-estimator is robust in the sense of having positive *finite-sample breakdown point* and uniformly bounded influence function, that is, finite gross error sensitivity (Hampel et al. (1986), Sections 2.2 and 2.3).

LEMMA 5. *In the setting of Lemma 1, define* $z_{\min} := \inf(\operatorname{supp} p_0)$ *and* $z_{\max} := \sup(\operatorname{supp} p_0)$. *Then*

(a) $\lim_{z \to -\infty} \psi_0^*(z) < \infty$ *if and only if* $\limsup_{z \searrow z_{\min}} h_0(z) < \infty$, *in which case* $z_{\min} = -\infty$;

(b) $\lim_{z \to \infty} \psi_0^*(z) > -\infty$ *if and only if* $\limsup_{z \nearrow z_{\max}} h_0(z) < \infty$, *in which case* $z_{\max} = \infty$.

Recall that a Laplace density has a constant two-sided hazard function, as well as a score function whose absolute value is constant. Roughly speaking, the conditions on h_0 in Lemma 5 are satisfied by densities whose tails are heavier than those of the Laplace density (Samworth and Johnson (2004)), for which it is particularly attractive to have bounded projected score functions.

2.2. *The log-concave Fisher divergence projection.* Let P_0 and P_1 be Borel probability measures on $\mathbb{R}$ such that P_0 is absolutely continuous with respect to P_1. Write $\operatorname{supp} P_0$ for the support of P_0 (i.e., the smallest closed set S satisfying $P_0(S) = 1$), and $\operatorname{Int}(\operatorname{supp} P_0)$ for its interior. Suppose that there exists a Radon–Nikodym derivative dP_0/dP_1 that is continuous on $\operatorname{Int}(\operatorname{supp} P_0)$, and also strictly positive and differentiable on some subset $E \subseteq \operatorname{Int}(\operatorname{supp} P_0)$ such that $P_0(E^c) = 0$.[4] The *Fisher divergence* (also known as the *Fisher information distance*[5]) from P_1 to P_0 is defined to be

$$I(P_0, P_1) := \int_E \left(\left(\log \frac{dP_0}{dP_1} \right)' \right)^2 dP_0. \tag{22}$$

[4]This condition precludes P_0 from having any isolated atoms, so in particular, P_0 cannot be a discrete measure.

[5]This is not to be confused with the *Fisher information* (or *Fisher–Rao*) *metric* (Amari and Nagaoka (2000), Chapter 2), a Riemannian metric on a manifold of probability distributions.

If P_0, P_1 do not satisfy the assumptions above, then we define $I(P_0, P_1) := \infty$. In the case where P_0, P_1 have Lebesgue densities p_0, p_1, respectively, which are both locally absolutely continuous on $\mathbb{R}$, we have

$$I(p_0, p_1) \equiv I(P_0, P_1) = \begin{cases} \displaystyle\int_{\{p_0>0\}} \left(\left(\log \frac{p_0}{p_1} \right)' \right)^2 p_0 & \text{if } \operatorname{supp} p_0 \subseteq \operatorname{supp} p_1, \\ \infty & \text{otherwise.} \end{cases}$$

For further background on the Fisher divergence, see Johnson (2004), Definition 1.13, Yang, Martin and Bondell (2019), Section 2 and references therein.

The following lemma establishes the connection between the projected score function and the Fisher divergence.

LEMMA 6. *In the setting of Lemma 1, there is a unique continuous log-concave density p_0^* on $\mathbb{R}$ such that* $\operatorname{supp} p_0^* = \mathcal{S}_0$ *and* $\log p_0^*$ *has right derivative ψ_0^* on $\mathcal{S}_0$. In particular, $\psi_0^* = (\log p_0^*)'$ Lebesgue almost everywhere on $\mathcal{S}_0$. Furthermore, if p_0 is absolutely continuous, then p_0^* minimises $I(p_0, p)$ over the class $\mathcal{P}_{\mathrm{LC}}$ of all univariate log-concave densities p, and if $p_0 \in \mathcal{P}_{\mathrm{LC}}$, then $p_0^* = p_0$.*

Even when p_0 is only uniformly continuous, we refer to p_0^* as the *log-concave Fisher divergence projection* of p_0. In contrast, the log-concave maximum likelihood projection p_0^{ML} of the distribution P_0 (Dümbgen, Samworth and Schuhmacher (2011), Barber and Samworth (2021)) can be interpreted as a minimiser of Kullback–Leibler divergence rather than Fisher divergence over the class of upper semicontinuous log-concave densities. By Dümbgen, Samworth and Schuhmacher (2011), Theorem 2.2, p_0^{ML} exists and is unique if and only if P_0 is nondegenerate and has a finite mean (but not necessarily a Lebesgue density). On the other hand, moment conditions are not required for p_0^* to exist and be unique, but p_0^* is only defined in Lemma 6 when P_0 has a uniformly continuous density on $\mathbb{R}$. As we will discuss in Section 3.1, the nonexistence of the Fisher divergence projection for discrete measures P_0 has consequences for our statistical methodology.

When p_0 is not log-concave, p_0^{ML} usually does not coincide with p_0^* even when both exist and, moreover, the associated regression M-estimators

$$\hat{\beta}_{\psi_0^{\mathrm{ML}}} \in \operatorname*{argmax}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \log p_0^{\mathrm{ML}}(Y_i - X_i^\top \beta) \quad \text{and} \quad \hat{\beta}_{\psi_0^*} \in \operatorname*{argmax}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \log p_0^*(Y_i - X_i^\top \beta) \tag{23}$$

are generally different; see Examples S8 and S10 in Section S2. In fact, the following result shows that there exist error distributions P_0 for which the asymptotic covariance of $\hat{\beta}_{\psi_0^{\mathrm{ML}}}$ is arbitrarily large compared with that of the optimal convex M-estimator $\hat{\beta}_{\psi_0^*}$, even when the latter is close to being asymptotically efficient in the sense of (4).

PROPOSITION 7. *For every $\epsilon \in (0, 1)$, there exists a distribution P_0 with a finite mean and an absolutely continuous density p_0 such that $i(p_0) < \infty$, and the log-concave maximum likelihood projection $q_0 \equiv p_0^{\mathrm{ML}}$ has corresponding score function $\psi_0^{\mathrm{ML}} := q_0^{(\mathrm{R})}/q_0 \in \Psi_\downarrow(p_0)$ satisfying*

$$\frac{V_{p_0}(\psi_0^*)}{V_{p_0}(\psi_0^{\mathrm{ML}})} \le \epsilon \quad \textit{and} \quad \mathrm{ARE}^*(p_0) \ge 1 - \epsilon. \tag{24}$$

The idea for the proof of Proposition 7 is to construct an absolutely continuous density p_0 whose score function is constant on each of $(-\infty, -1)$, $(-1, 0)$, $(0, 1)$ and $(1, \infty)$; see Figure 3. The key point is that the log-concave maximum likelihood projection is constant on

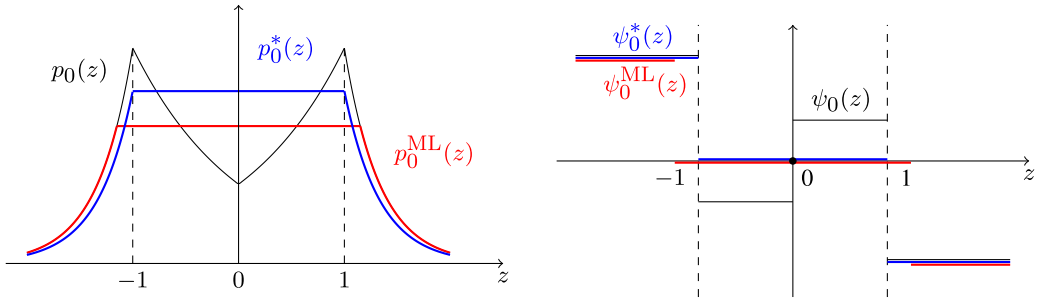

FIG. 3. *Illustration of the construction in the proof of Proposition 7. Left*: *Plot of the density p_0 (black) together with its log-concave maximum likelihood projection p_0^{ML} (red) and Fisher divergence projection p_0^* (blue). Right*: *Plot of the corresponding score functions.*

$[-1, 1]$ and exactly matches the true density in the tails, so in order for p_0^{ML} to integrate to 1, the densities and score functions can only agree on $(-\infty, -(1+\delta))$ and $(1+\delta, \infty)$ for some $\delta > 0$. On the other hand, the log-concave Fisher divergence projection matches the score ψ_0 on the whole of $[-1, 1]^c$ by merely being *proportional* to the true density in the tails. The main contribution to the Fisher information of p_0 arises from the region $(-(1+\delta), -1) \cup (1, 1+\delta)$, on which the antitonic score function approximation is exact but ψ_0^{ML} is equal to 0. This means that the antitonic relative efficiency is close to 1, while the log-concave maximum likelihood projection incurs a considerable relative efficiency loss over this region.

Our next result provides a simple lower bound on the antitonic information.

LEMMA 8. *Suppose that p_0 is an absolutely continuous density on $\mathbb{R}$ with $i(p_0) < \infty$. Then p_0 is bounded with $i^*(p_0) \geq 4\|p_0\|_\infty^2$, so*

$$\mathrm{ARE}^*(p_0) \geq \frac{4\|p_0\|_\infty^2}{i(p_0)}.$$

Equality holds if and only if p_0^* is a Laplace density, that is, there exist $\mu \in \mathbb{R}$ and $\sigma > 0$ such that $p_0^*(z) = (2\sigma)^{-1}\exp(-|z-\mu|/\sigma)$ for all $z \in \mathbb{R}$.

REMARK 9. A reassuring property of the antitonic projection is its affine equivariance: if p_0 is a uniformly continuous density, then for $a > 0$ and $b \in \mathbb{R}$, the density $z \mapsto a p_0(az+b) =: p_{a,b}(z)$ has antitonic projected score function and log-concave Fisher divergence projection given by

$$\psi_{a,b}^*(z) := a\psi_0^*(az+b) \quad \text{and} \quad p_{a,b}^*(z) := a p_0^*(az+b),$$

respectively, for $z \in \mathbb{R}$. It follows that $1/V_{p_{a,b}}(\psi_{a,b}^*) = i^*(p_{a,b}) = \int_{\mathbb{R}}(\psi_{a,b}^*)^2 p_{a,b} = a^2 i^*(p_0)$, so because $\|p_{a,b}\|_\infty = a\|p_0\|_\infty$, both the antitonic relative efficiency and the lower bound in Lemma 8 are affine invariant in the sense that they remain unchanged if we replace p_0 with $p_{a,b}$.

Similarly, if P_0 has a finite mean, then by the affine equivariance of the log-concave maximum likelihood projection (Dümbgen, Samworth and Schuhmacher (2011), Remark 2.4), $p_{a,b}^{\mathrm{ML}}(z) = a p_0^{\mathrm{ML}}(az+b)$, and hence $\psi_{a,b}^{\mathrm{ML}} = a\psi_0^{\mathrm{ML}}(az+b)$ for $z \in \mathbb{R}$. Thus, $V_{p_{a,b}}(\psi_{a,b}^{\mathrm{ML}}) = V_{p_0}(\psi_0^{\mathrm{ML}})/a^2$, so the first ratio in (24) is also affine invariant. Consequently, for any $C \in (0, \infty)$, $\mu \in \mathbb{R}$ and $\epsilon \in (0, 1)$, there exists a density p_0 satisfying (24) with $i(p_0) = C$ and $\int_{\mathbb{R}} z p_0(z)\, dz = \mu$.

The final result in this subsection relates properties of densities and their log-concave Fisher divergence projections.

PROPOSITION 10. *For a uniformly continuous density $p_0\colon \mathbb{R} \to \mathbb{R}$, the log-concave Fisher divergence projection p_0^* and its corresponding distribution function F_0^* : $[-\infty, \infty] \to \mathbb{R}$ have the following properties*:

(a) *Denote by $\mathcal{T}$ the set of $z \in \mathcal{S}_0$ such that ψ_0^* is nonconstant on every open interval containing z. For $z \in \mathcal{T}$, we have*

$$\frac{p_0^*(z)}{F_0^*(z)} \le \frac{p_0(z)}{F_0(z)} \quad \text{and} \quad \frac{p_0^*(z)}{1 - F_0^*(z)} \le \frac{p_0(z)}{1 - F_0(z)}, \tag{25}$$

whence $p_0^*(z) \le p_0(z)$.

(b) *$\|p_0^*\|_\infty \le \|p_0\|_\infty$ and $i(p_0^*) = -\int_{\mathbb{R}} p_0^* \, d\psi_0^* \le -\int_{\mathbb{R}} p_0 \, d\psi_0^* = i^*(p_0)$.*

We can define the two-sided hazard function h_0^* of the log-concave Fisher divergence projection p_0^* analogously to h_0 in (21). Since p_0^* is log-concave, its density quantile function $J_0^* := p_0^* \circ (F_0^*)^{-1}$ has decreasing right derivative $(J_0^*)^{(\mathrm{R})} = \psi_0^* \circ (F_0^*)^{-1}$ by (the proof of) Lemma 6, so J_0^* is concave on $[0, 1]$. Thus,

$$0 = J_0^*(0) \le J_0^*(u) - u\big(J_0^*\big)^{(\mathrm{R})}(u) \quad \text{and} \quad 0 = J_0^*(1) \le J_0^*(u) + (1-u)\big(J_0^*\big)^{(\mathrm{R})}(u)$$

for all $u \in [0, 1]$, so for $z \in \mathcal{T}$, Lemma 1 and (25) in Proposition 10*(a)* imply that

$$\begin{aligned} |\psi_0^*(z)| = \big|\big(J_0^*\big)^{(\mathrm{R})}\big(F_0^*(z)\big)\big| &\le \frac{J_0^*(F_0^*(z))}{F_0^*(z) \wedge (1 - F_0^*(z))} = \frac{p_0^*(z)}{F_0^*(z) \wedge (1 - F_0^*(z))} = h_0^*(z) \\ &\le \frac{p_0(z)}{F_0(z) \wedge (1 - F_0(z))} = h_0(z). \end{aligned}$$

Proposition 10*(b)* provides inequalities on the supremum norm and antitonic information of the log-concave Fisher divergence projection. In particular, the Fisher information of the projected density is at most the antitonic information of the original density.

2.3. *Examples*.

EXAMPLE 11. First, let p_0 be the Beta(a, b) density given by $p_0(z) = z^{a-1}(1 - z)^{b-1}\mathbb{1}_{\{z\in(0,1)\}}/\mathrm{B}(a, b)$ for $a, b > 1$, where B denotes the beta function. Then p_0 is uniformly continuous and log-concave on $\mathbb{R}$, so

$$\psi_0^*(z) = \psi_0(z) = (\log p_0)'(z) = \frac{a-1}{z} - \frac{b-1}{1-z}$$

for all $z \in (0, 1)$, while $\psi_0^*(z) = \infty$ for $z \le 0$ and $\psi_0^*(z) = -\infty$ for $z \ge 1$. We have $i^*(p_0) = i(p_0) = \int_0^1 \psi_0^2 p_0 < \infty$ if and only if $a, b > 2$, in which case the conclusions of Theorem 2*(c, d)* hold.

EXAMPLE 12. Let p_0 be the standard Cauchy density given by $p_0(z) = 1/(\pi(1 + z^2))$ for $z \in \mathbb{R}$. Then p_0 is absolutely continuous on $\mathbb{R}$ with $i(p_0) = \int_{\mathbb{R}} (p_0')^2/p_0 = 1/2$ and $\lim_{z\to\pm\infty} h_0(z) = 0$. We will derive an explicit expression for ψ_0^*, which is necessarily

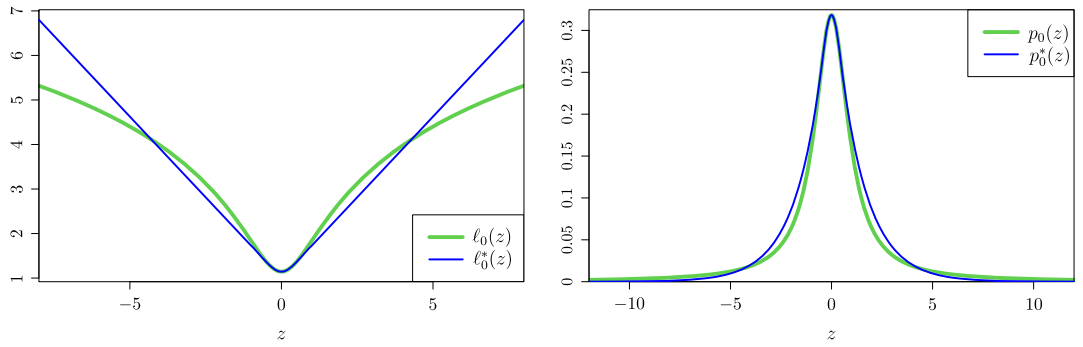

FIG. 4. Left: *The negative log-density* $\ell_0 = -\log p_0$ *and the optimal convex loss function* $\ell_0^* = -\log p_0^*$ *when* p_0 *is the standard Cauchy density.* Right: *The corresponding densities* p_0 *and* p_0^*.

bounded by Lemma 5. In contrast to the previous example, p_0 is not log-concave, so ψ_0^* does not coincide with $\psi_0 = p_0'/p_0 \colon z \mapsto -2z/(1+z^2)$. Indeed,

$$F_0(z) = \frac{1}{2} + \frac{\arctan(z)}{\pi} \quad \text{for } z \in \mathbb{R},$$

$$F_0^{-1}(u) = \tan\left(\pi\left(u - \frac{1}{2}\right)\right) = -\cot(\pi u) \quad \text{for } u \in (0,1),$$

$$J_0(u) = \frac{1}{\pi(1+\cot^2(\pi u))} = \frac{\sin^2(\pi u)}{\pi} = \frac{1-\cos(2\pi u)}{2\pi} \quad \text{for } u \in [0,1].$$

Let $t_0 \approx 2.33$ be the unique $t \in (0,\pi)$ satisfying $t = \tan(t/2)$, and define $u_0 := t_0/(2\pi) \in (0, 1/2)$. Then we can verify that $\hat{J}_0$ is linear on $[0, u_0]$ and on $[1-u_0, 1]$, with $\hat{J}_0(u) = J_0(u)$ for $u \in [u_0, 1-u_0] \cup \{0,1\}$; see the left panel of Figure 2. It follows that

$$\hat{J}_0^{(\mathrm{R})}(u) = \begin{cases} \sin t_0 & \text{for } u \in [0, u_0], \\ \sin(2\pi u) & \text{for } u \in [u_0, 1-u_0], \\ -\sin t_0 & \text{for } u \in [1-u_0, 1]; \end{cases}$$

$$\psi_0^*(z) = \hat{J}_0^{(\mathrm{R})}(F_0(z)) = \begin{cases} \sin t_0 = -2z_0/(1+z_0^2) & \text{for } z \in (-\infty, -z_0], \\ -\sin(2\arctan z) = -2z/(1+z^2) = \psi_0(z) & \text{for } z \in [-z_0, z_0], \\ -\sin t_0 & \text{for } z \in [z_0, \infty), \end{cases}$$

where $z_0 := \cot(\pi u_0) = \cot(t_0/2) \approx 0.43 \in (0,1)$ satisfies $z_0 \arctan(1/z_0) = 1/2$. Thus, $\psi_0^*(z) = \psi_0((z \wedge z_0) \vee (-z_0))$ for $z \in \mathbb{R}$; see the right panel of Figure 2. An antiderivative ϕ_0^* of ψ_0^* is given by

$$\phi_0^*(z) := \int_0^z \psi_0^* - \log\pi = \begin{cases} -\log(\pi(1+z^2)) & \text{for } z \in [-z_0, z_0], \\ -(|z| - z_0)\sin t_0 - \log(\pi(1+z_0^2)) & \text{for } z \in \mathbb{R} \setminus [-z_0, z_0]. \end{cases}$$

As illustrated in the left panel of Figure 4, $\ell_0^* := -\phi_0^*$ is a symmetric convex function that is approximately quadratic on $[-z_0, z_0]$ and linear outside this interval, so in this respect, it resembles the Huber loss function (7). This is significant as far as M-estimation is concerned, since Huber-like loss functions are designed precisely to be robust to outliers, such as those that arise in regression problems with heavy-tailed Cauchy errors. As discussed in the introduction, ℓ_0^* is optimal in the sense that the resulting regression M-estimator $\hat{\beta}_{\psi_0^*} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \ell_0^*(Y_i - X_i^\top \beta)$ has minimal asymptotic covariance among all convex

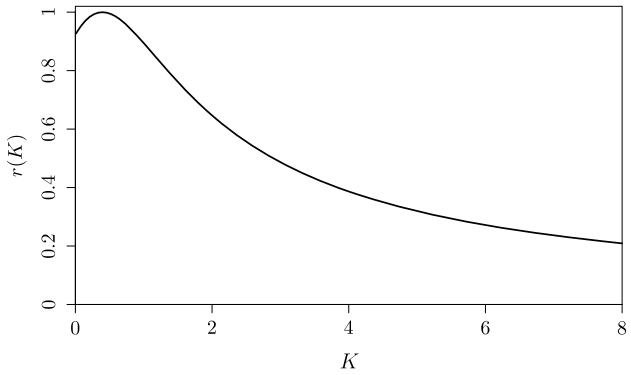

FIG. 5. *Plot of the asymptotic relative efficiency $r(K)$ of the Huber M-estimator $\hat{\beta}_{\psi_K}$ compared with the optimal convex M-estimator.*

M-estimators. By direct computation,

$$\frac{1}{V_{p_0}(\psi_0^*)} = i^*(p_0) = \frac{1}{2} - \frac{2t_0\cos(2t_0) - \sin(2t_0)}{4\pi} \approx 0.439,$$

$$\text{so } \mathrm{ARE}^*(p_0) = \frac{i^*(p_0)}{i(p_0)} \approx 0.878$$

in this case, meaning that the restriction to convex loss functions results in only a small loss of efficiency relative to the maximum likelihood estimator. This may well be outweighed by the increased computational convenience of optimising a convex empirical risk function as opposed to a Cauchy likelihood function, which typically has several local extrema; see van der Vaart (1998), Example 5.50, for a discussion of the difficulties involved. Lemma 8 yields the bound $\mathrm{ARE}^*(p_0) \geq 8\|p_0\|_\infty^2 = 8/\pi^2 \approx 0.811$.

For $K > 0$, the Huber regression M-estimator $\hat{\beta}_{\psi_K} \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \sum_{i=1}^n \ell_K(Y_i - X_i^\top\beta)$ defined with respect to (7) has asymptotic relative efficiency

$$r(K) := \frac{V_{p_0}(\psi_0^*)}{V_{p_0}(\psi_K)} = \frac{4\arctan^2 K}{\pi(\pi K^2 + 2K - 2(1+K^2)\arctan K)i^*(p_0)}$$

compared with the optimal convex M-estimator $\hat{\beta}_{\psi_0^*}$; see Figure 5. The maximum value $\sup_{K>0} r(K) \approx 0.9998$ is attained at $K^* \approx 0.394$. Moreover,

$$\lim_{K\to 0} r(K) = \frac{4}{\pi^2 i^*(p_0)} \approx 0.922$$

is the asymptotic relative efficiency $V_{p_0}(\psi_0^*)/V_{p_0}(\psi)$ of the least absolute deviation (LAD) estimator $\hat{\beta}_\psi \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \sum_{i=1}^n |Y_i - X_i^\top\beta|$, for which $\psi(\cdot) := -\operatorname{sgn}(\cdot)$ and $V_{p_0}(\psi) = 1/(4p_0(0)^2) = \pi^2/4$. On the other hand, the Huber loss ℓ_K in (7) converges pointwise to the squared error loss as $K \to \infty$, so $\lim_{K\to\infty} V_{p_0}(\psi_K) = \int_{\mathbb{R}} z^2 p_0(z)\,dz = \infty$, and hence $\lim_{K\to\infty} r(K) = 0$. Recall from the discussion in the introduction the difficulties of choosing K, and its connection to the choice of scale.

The Cauchy density p_0 and its log-concave Fisher divergence projection $p_0^* := e^{\phi_0^*}$ are plotted in the right panel of Figure 4. Since $p_0 = p_0^*$ on $[-z_0, z_0]$ and ψ_0^* is constant on $\mathbb{R}\setminus[-z_0, z_0]$, it turns out that $i^*(p_0) = -\int_{\mathbb{R}} p_0\,d\psi_0^* = -\int_{\mathbb{R}} p_0^*\,d\psi_0^* = i(p_0^*)$; so, both inequalities in Proposition 10*(b)* are in fact equalities in this example.

Section S2 characterises the antitonic score projection for a variety of other densities p_0. In Example S8, we take P_0 to be a scaled t_2 distribution, which has a finite first moment (unlike

the Cauchy distribution in Example 12), and verify that $\hat{\beta}_{\psi_0^{\mathrm{ML}}}$ and $\hat{\beta}_{\psi_0^*}$ in (23) are different convex M-estimators. Example S9 features a symmetrised Pareto density with polynomially decaying tails, where the optimal convex loss function ℓ_0^* is a scale transformation of the robust absolute error loss $z \mapsto |z|$. Moving on from heavy-tailed distributions, Example S10 considers a density p_0 that fails to be log-concave because it is not unimodal, while Proposition S11 is a general result about the log-concave maximum likelihood and Fisher divergence projections of Gaussian mixtures.

3. Semiparametric M-estimation via antitonic score matching.

3.1. *Warm-up*: *Estimation of the projected score function from direct observations*. Section 2 was concerned with the properties of the antitonic score projection on the population level. Ultimately, our goal is to be able to incorporate these insights into a linear regression setting, but in this subsection we address an intermediate aim. Specifically, we consider the nonparametric estimation of the antitonic projected score function ψ_0^* based on a sample $\varepsilon_1, \dots, \varepsilon_n \overset{\mathrm{iid}}{\sim} p_0$. This is an interesting problem in its own right, but an idealised one for the purposes of the regression setting that we have in mind, since there we do not observe the regression errors $\varepsilon_1, \dots, \varepsilon_n$ directly, and instead will need to rely on residuals from a pilot fit as proxies.

We first explain why a naive approach to antitonic score matching fails before describing our alternative solution. For a locally absolutely continuous function $\psi : \mathbb{R} \to \mathbb{R}$ with derivative ψ', the empirical analogue of $D_{p_0}(\psi) = \mathbb{E}(\psi^2(\varepsilon_1) + 2\psi'(\varepsilon_1))$ in (14) is

$$\hat{D}_n(\psi) \equiv \hat{D}_n(\psi; \varepsilon_1, \dots, \varepsilon_n) := \frac{1}{n} \sum_{i=1}^{n} \{\psi^2(\varepsilon_i) + 2\psi'(\varepsilon_i)\}. \tag{26}$$

Recall from the introduction that score matching estimates a score function by an empirical risk minimiser $\hat{\psi}_n \in \operatorname{argmin}_{\psi \in \Psi} \hat{D}_n(\psi)$ over an appropriate class of functions Ψ.

However, to obtain a monotone score estimate, we cannot minimise $\psi \mapsto \hat{D}_n(\psi)$ directly over the class $\Psi_{\downarrow}^{\mathrm{ac}}$ of all decreasing, locally absolutely continuous $\psi : \mathbb{R} \to \mathbb{R}$. Indeed, $\inf_{\psi \in \Psi_{\downarrow}^{\mathrm{ac}}} \hat{D}_n(\psi) = -\infty$, as can be seen by constructing differentiable approximations to a decreasing step function whose jumps are at the data points $\varepsilon_1, \dots, \varepsilon_n$. To circumvent this issue, we instead propose the following estimation strategy.

Antitonic projected score estimation: Consider smoothing the empirical distribution of $\varepsilon_1, \dots, \varepsilon_n$, for example, by convolving it with an absolutely continuous kernel $K : \mathbb{R} \to \mathbb{R}$ to obtain a kernel density estimator $z \mapsto \tilde{p}_n(z) := n^{-1} \sum_{i=1}^{n} K_h(z - \varepsilon_i)$, where $h > 0$ is a suitable bandwidth and $K_h(\cdot) := h^{-1} K(\cdot/h)$. We can then define the smoothed empirical score matching objective

$$\tilde{D}_n(\psi) := D_{\tilde{p}_n}(\psi) = \int_{\mathbb{R}} \psi^2 \tilde{p}_n + 2 \int_{\mathcal{S}_0} \tilde{p}_n \, d\psi$$

for $\psi \in \Psi_{\downarrow}(\tilde{p}_n)$, which approximates the population expectation in the definition of $D_{p_0}(\psi)$. By Theorem 2,

$$\hat{J}_n^{(\mathrm{R})} \circ \tilde{F}_n \in \operatorname*{argmin}_{\psi \in \Psi_{\downarrow}(\tilde{p}_n)} \tilde{D}_n(\psi), \tag{27}$$

where $\tilde{F}_n$ denotes the distribution function corresponding to $\tilde{p}_n$, and $J_n := \tilde{p}_n \circ \tilde{F}_n^{-1}$. This is reminiscent of maximum smoothed likelihood estimation of a density or distribution function (e.g., Eggermont and LaRiccia (2000); Groeneboom and Jongbloed (2014), Sections 8.2 and 8.5). By evaluating the antiderivative of J_n on a suitably fine grid, we may obtain a

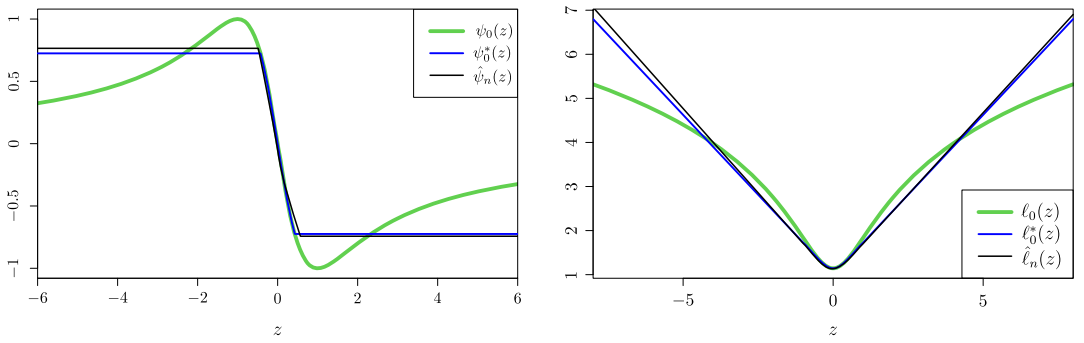

FIG. 6. *Kernel-based estimates of the projected score function and optimal convex loss function based on a sample of size* $n = 2000$ *from the Cauchy distribution.*

piecewise affine approximation to $\hat{J}_n$ (and hence a piecewise constant approximation to its derivative) using PAVA, whose space and time complexities scale linearly with the size of the grid (Samworth and Shah (2025), Section 9.3.1).

More generally, we can use $\varepsilon_1, \ldots, \varepsilon_n$ to construct a generic (not necessarily monotone) score estimator $\tilde{\psi}_n$ and an estimate $\hat{F}_n$ of the distribution function F_0 corresponding to the density p_0. By analogy with the explicit representation (17) of ψ_0^*, we then define the decreasing score estimate

$$\hat{\psi}_n := \widehat{\mathcal{M}}_{\mathrm{R}}(\tilde{\psi}_n \circ \hat{F}_n^{-1}) \circ \hat{F}_n. \tag{28}$$

As explained above, (a numerical approximation to) $\hat{\psi}_n$ can be computed efficiently using isotonic regression algorithms. In particular, if $\hat{F}_n$ is taken to be the empirical distribution function of $\varepsilon_1, \ldots, \varepsilon_n$, then by Proposition S33, $\hat{\psi}_n^{(\mathrm{L})} := \widehat{\mathcal{M}}_{\mathrm{L}}(\tilde{\psi}_n \circ \hat{F}_n^{-1}) \circ \hat{F}_n$ is an antitonic least squares estimator based on $\{(\varepsilon_i, \tilde{\psi}_n(\varepsilon_i)) : i \in [n]\}$. Our decreasing score estimate can be taken to be either $\hat{\psi}_n^{(\mathrm{L})}$ or the closely related $\hat{\psi}_n$: for every $z \in \mathbb{R}$, we have $\hat{\psi}_n(z) = \hat{\psi}_n^{(\mathrm{L})}(\varepsilon(z))$, where $\varepsilon(z)$ is the smallest element of $\{\varepsilon_1, \ldots, \varepsilon_n\}$ that is strictly greater than z (where such an element exists) or equal to $\max_{i \in [n]} \varepsilon_i$ otherwise.

The transformation (28) may be applied to any appropriate initial score estimator $\tilde{\psi}_n$. Since a misspecified parametric method may introduce significant error at the outset, we seek a nonparametric estimator. For instance, we may take $\tilde{\psi}_n$ to be a ratio of kernel density estimates of p_0' and p_0, which may be truncated for theoretical and practical convenience to avoid instability in low-density regions; see (31) in Section 3.3. Observe that (27) is a special case of (28) with $\tilde{\psi}_n = \tilde{p}_n'/\tilde{p}_n$ and $\hat{F}_n = \tilde{F}_n$. Figures 6 and 7 illustrate the kernel-based

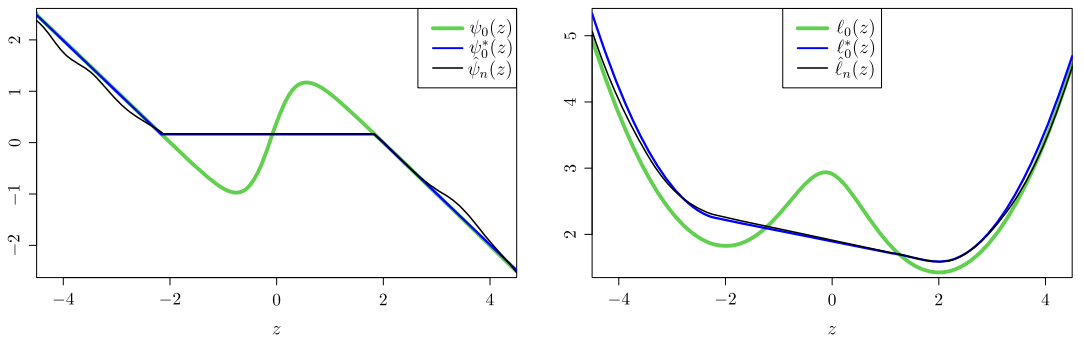

FIG. 7. *Kernel-based estimates of the projected score function and optimal convex loss function based on a sample of size* $n = 10^4$ *from the Gaussian mixture distribution* $0.4N(-2, 1) + 0.6N(2, 1)$.

projected score estimate $\hat{\psi}_n$ and its negative antiderivative $\hat{\ell}_n$ (subsequently used as a data-driven proxy for the optimal convex loss ℓ_0^*) based on samples from Cauchy and Gaussian mixture distributions respectively.

3.2. *Linear regression*: *Alternating algorithm outline*. Suppose that we observe independent and identically distributed pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ satisfying

$$Y_i = X_i^\top \beta_0 + \varepsilon_i \tag{29}$$

for $i \in [n]$, where $X_1, \ldots, X_n$ are $\mathbb{R}^d$-valued covariates that are independent of errors $\varepsilon_1, \ldots, \varepsilon_n$ with an unknown absolutely continuous (Lebesgue) density p_0 on $\mathbb{R}$. A necessary condition for β_0 to be identifiable is that $\mathbb{E}(X_1 X_1^\top)$ is positive definite, since otherwise there exists $v \in \mathbb{R}^d \setminus \{0\}$ such that $X_i^\top v = 0$ for all i almost surely. In this case, the joint distribution of our observed data is unchanged if we replace β_0 with $\beta_0 + v$. To accommodate heavy-tailed (e.g., Cauchy) errors, we do not necessarily insist that $\mathbb{E}(\varepsilon_1) = 0$, or even suppose that ε_1 is integrable, though see also the discussion at the start of Section 3.4.

As the regression coefficients β_0, the error density p_0, and hence the antitonic score projected score ψ_0^* are unknown, a natural estimation strategy on the population level is to alternate between the following two steps:

I. For a fixed β, minimise the (convex) score matching objective $D_{q_\beta}(\psi)$ based on the density q_β of $Y_1 - X_1^\top \beta$.

II. For a fixed decreasing and right-continuous ψ, minimise the convex function $\beta \mapsto \mathbb{E}\ell(Y_1 - X_1^\top \beta)$, where ℓ is a negative antiderivative of ψ.

This approach can be motivated by a joint optimisation problem over a set of pairs (β, ψ); see Section S3.1. In an empirical version of this alternating algorithm based on $(X_1, Y_1), \ldots, (X_n, Y_n)$, we can estimate ψ_0^* in Step I by minimising a sample analogue of $D_{q_\beta}(\psi)$ over ψ. As discussed in Section 3.1, the unknown density of $Y_1 - X_1^\top \beta$ can be approximated by smoothing the empirical distribution of the residuals $(Y_i - X_i^\top \beta)_{i=1}^n$ from the current estimate of β_0. Step II then involves finding an M-estimator (1) of β_0 based on the convex loss function induced by the current estimate of ψ_0^*. In practice, we can apply any suitable convex optimisation algorithm such as gradient descent, with Newton's method being a faster alternative when the score estimate is differentiable. Steps I and II can then be iterated to convergence.

The following two subsections focus in turn on linear regression with symmetric errors and with an explicit intercept term. We will analyse a specific version of the above procedure that is initialised with a pilot estimator $\bar{\beta}_n$ of β_0. Provided that $(\bar{\beta}_n)$ is $\sqrt{n}$-consistent, we show that a single iteration of Steps I and II yields a semiparametric convex M-estimator of β_0 that achieves "antitonic efficiency" as $n \to \infty$.

3.3. *Linear regression with symmetric errors*. Under the assumption that p_0 is symmetric, we first approximate ψ_0^* via antitonic projection of kernel-based score estimators. We do not observe the errors $\varepsilon_1, \ldots, \varepsilon_n$ directly, so in view of the discussion above, in the algorithm below we use the residuals from the pilot regression estimator to construct our initial score estimators. Assume throughout that $n \geq 3$.

1. *Sample splitting*: Partition the observations into three folds indexed by disjoint subsets $I_1, I_2, I_3 \subseteq [n]$ such that $|I_1| = |I_2| = \lfloor n/3 \rfloor$ and $|I_3| = n - |I_1| - |I_2|$, respectively. For notational convenience, let $I_{j+3} := I_j$ for $j \in \{1, 2\}$.

2. *Pilot estimators*: Fix a convex function $L\colon \mathbb{R} \to \mathbb{R}$, and for $j \in \{1,2,3\}$, let $\bar{\beta}_n^{(j)}$ be an M-estimator

$$\bar{\beta}_n^{(j)} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \sum_{i\in I_j} L(Y_i - X_i^\top \beta). \tag{30}$$

3. *Antitonic projected score estimation*: For $j \in \{1,2,3\}$ and $i \in I_{j+1}$, define out-of-sample residuals $\hat{\varepsilon}_i := Y_i - X_i^\top \bar{\beta}_n^{(j)}$. Letting $K\colon \mathbb{R} \to [0,\infty)$ be a differentiable kernel and $h \equiv h_n > 0$ be a (deterministic) bandwidth, define a kernel density estimator $\tilde{p}_{n,j}$ of p_0 by

$$\tilde{p}_{n,j}(z) := \frac{1}{|I_{j+1}|} \sum_{i\in I_{j+1}} K_h(z - \hat{\varepsilon}_i)$$

for $z \in \mathbb{R}$, where $K_h(\cdot) = h^{-1}K(\cdot/h)$. In addition, let $\tilde{S}_{n,j} := \{z \in \mathbb{R} : |\tilde{p}'_{n,j}(z)| \le \alpha_n, \tilde{p}_{n,j}(z) \ge \gamma_n\}$, where $\alpha_n \in (0,\infty]$ and $\gamma_n \in (0,\infty)$ are truncation parameters, and define $\tilde{\psi}_{n,j}\colon \mathbb{R} \to \mathbb{R}$ by

$$\tilde{\psi}_{n,j}(z) := \frac{\tilde{p}'_{n,j}(z)}{\tilde{p}_{n,j}(z)} \mathbb{1}_{\{z\in\tilde{S}_{n,j}\}}. \tag{31}$$

Writing $\tilde{F}_{n,j}$ for the distribution function corresponding to $\tilde{p}_{n,j}$, let $\hat{\psi}_{n,j} := \widehat{\mathcal{M}}_{\mathrm{R}}(\tilde{\psi}_{n,j} \circ \tilde{F}_{n,j}^{-1}) \circ \tilde{F}_{n,j}$ be an antitonic projected score estimate, in accordance with (28). Finally, define an estimator $\hat{\psi}_{n,j}^{\mathrm{anti}} \in \Psi_{\downarrow}^{\mathrm{anti}}(p_0)$ of ψ_0^* by

$$\hat{\psi}_{n,j}^{\mathrm{anti}}(z) := \frac{\hat{\psi}_{n,j}(z) - \hat{\psi}_{n,j}(-z)}{2}$$

for $z \in \mathbb{R}$.

4. *Plug-in cross-fitted convex M-estimator*: For $j \in \{1,2,3\}$, let $\hat{\ell}_{n,j}^{\mathrm{sym}}\colon \mathbb{R} \to \mathbb{R}$ be the induced convex loss function given by $\hat{\ell}_{n,j}^{\mathrm{sym}}(z) := -\int_0^z \hat{\psi}_{n,j}^{\mathrm{anti}}$, and define

$$\hat{\beta}_n^{(j)} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \sum_{i\in I_{j+2}} \hat{\ell}_{n,j}^{\mathrm{sym}}(Y_i - X_i^\top \beta) \tag{32}$$

to be a corresponding M-estimator of β_0. Finally, let

$$\hat{\beta}_n^{\dagger} := \frac{\hat{\beta}_n^{(1)} + \hat{\beta}_n^{(2)} + \hat{\beta}_n^{(3)}}{3}. \tag{33}$$

In (32), $\hat{\beta}_n^{(j)}$ always exists because either $\hat{\psi}_{n,j}^{\mathrm{anti}} \equiv 0$ and any $\beta \in \mathbb{R}^d$ is a minimiser, or otherwise $\inf_{z\in\mathbb{R}} \hat{\psi}_{n,j}^{\mathrm{anti}}(z) < 0 < \sup_{z\in\mathbb{R}} \hat{\psi}_{n,j}^{\mathrm{anti}}(z)$, and hence the convex function $\hat{\ell}_{n,j}^{\mathrm{sym}}$ is *coercive* in the sense that $\hat{\ell}_{n,j}^{\mathrm{sym}}(z) \to \infty$ as $|z| \to \infty$. Moreover, $\hat{\beta}_n^{(j)}$ is unique if $\hat{\psi}_{n,j}^{\mathrm{anti}}$ is strictly decreasing and the design matrix X has full column rank, which happens with probability tending to 1 as $n \to \infty$ if $\mathbb{E}(X_1 X_1^\top)$ is invertible; see Proposition S24*(a)* for elementary justifications of these claims. In practice, our antitonic score estimates may have constant pieces, but the conclusion of Theorem 13 below applies to all sequences of minimisers $(\hat{\beta}_n^{(j)})$, so is unaffected by nonuniqueness issues.

The above procedure uses the observations indexed by I_1, I_2, I_3 to construct the pilot estimator $\bar{\beta}_n^{(1)}$, antitonic score estimate $\hat{\psi}_{n,1}^{\mathrm{anti}}$ and semiparametric M-estimator $\hat{\beta}_n^{(1)}$ respectively. Since each fold (specifically the last one) contains only about one-third of all the data, sample splitting reduces the efficiency of $\hat{\beta}_n^{(1)}$. We remedy this by *cross-fitting* (van der Vaart (1998), page 393; Chernozhukov et al. (2018), Section 3), which involves cyclically permuting the

folds to obtain $\hat{\beta}_n^{(2)}, \hat{\beta}_n^{(3)}$ analogously to $\hat{\beta}_n^{(1)}$, and then averaging these three estimators. This reduces the limiting covariance of $\hat{\beta}_n^{(1)}$ by a factor of three in the theory below, where we show that $\hat{\beta}_n^{(1)}, \hat{\beta}_n^{(2)}, \hat{\beta}_n^{(3)}$ are "asymptotically independent" in a precise sense.

A different version of cross-fitting (Chernozhukov et al. (2018), Definition 3.2) instead averages the empirical risk functions across all three folds, and outputs a single estimator

$$\hat{\beta}_n^{\ddagger} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \sum_{j=1}^{3} \sum_{i \in I_{j+2}} \hat{\ell}_{n,j}^{\mathrm{sym}}(Y_i - X_i^\top \beta), \tag{34}$$

whose existence is similarly guaranteed by the convexity of $\hat{\ell}_{n,j}^{\mathrm{sym}}$ for $j \in \{1, 2, 3\}$.

To introduce our theoretical guarantees for this procedure, for a sequence of regression models (29) indexed by $n \in \mathbb{N}$, we make the following assumptions on the model and the parameters in our procedure.

(A1) p_0 is an absolutely continuous density on $\mathbb{R}$ such that $i(p_0) < \infty$ and $\int_{\mathbb{R}} |z|^\delta p_0(z)\, dz < \infty$ for some $\delta > 0$.

(A2) There exists $t_0 > 0$ such that for $t \in \{-t_0, t_0\}$, the antitonic projected score function ψ_0^* satisfies $\int_{\mathbb{R}} \psi_0^*(z+t)^2 p_0(z)\, dz < \infty$.

(A3) The kernel K is nonnegative, twice continuously differentiable and supported on $[-1, 1]$.

(A4) $\alpha_n \to \infty$, $\gamma_n, h_n \to 0$, $n h_n^3 \gamma_n^2 \to \infty$ and $(h_n \vee n^{-2\rho/3})(\alpha_n/\gamma_n)^2 \to 0$ for some $\rho \in (0, \delta/(\delta+1))$.

(A5) $\mathbb{E}(X_1 X_1^\top) \in \mathbb{R}^{d \times d}$ is positive definite and $\max_{i \in [n]} \|X_i\| \alpha_n/\gamma_n = o_p(n^{1/2})$ as $n \to \infty$.

The conditions (A1) and (A2) are satisfied by a wide variety of commonly encountered densities p_0 ranging from all t_ν densities with $\nu > 0$ degrees of freedom (including the Cauchy density as a special case $\nu = 1$) to lighter-tailed Weibull, Laplace, Gaussian and Gumbel densities. In particular, P_0 need not have a finite mean, and our procedure does not require knowledge of the exponent $\delta > 0$ in (A1). By Lemma 1, $\{z \in \mathbb{R} : \psi_0^*(z) \in \mathbb{R}\} = \mathcal{S}_0 = (\inf(\operatorname{supp} p_0), \sup(\operatorname{supp} p_0))$, so if (A2) holds, then $\mathcal{S}_0 = \mathbb{R}$ and ψ_0^* must be finite-valued on $\mathbb{R}$. The truncation parameters α_n, γ_n and bandwidth h_n can be chosen quite flexibly; for instance, (A4) holds if $\alpha_n = \gamma_n^{-1} = \log n$ and $h_n = n^{-b}$ for some $b \in (0, 1/3)$. As for (A5), the fact that $\|X_1\|^2, \ldots, \|X_n\|^2$ are identically distributed and integrable means that $\max_{i \in [n]} \|X_i\| = o_p(n^{1/2})$; see (S66). Our condition is slightly stronger than this to account for the score estimators being uniformly bounded in absolute value by α_n/γ_n. In particular, if $\mathbb{E}(\|X_1\|^3) < \infty$, then $\max_{i \in [n]} \|X_i\| = o_p(n^{1/3})$, so under (A4), we have $\alpha_n/\gamma_n = o_p(h_n^{-1/2}) = o_p(n^{1/6})$, and hence (A5) holds automatically.

We also require the pilot estimators in Step 2 to exist and be $\sqrt{n}$-consistent for β_0. By a classical result (see Proposition S24), this is guaranteed if the loss function L in this step has negative right derivative φ satisfying both $\mathbb{E}\varphi(\varepsilon_1) = 0$ and the following condition.

(B) φ is decreasing and right continuous with $\inf_{z \in \mathbb{R}} \varphi(z) < 0 < \sup_{z \in \mathbb{R}} \varphi(z)$. In addition,

$$V_{p_0}(\varphi) = \frac{\int_{\mathbb{R}} \varphi^2 p_0}{(\int_{\mathbb{R}} p_0\, d\varphi)^2} \in (0, \infty)$$

and there exists $t_0 > 0$ such that $\int_{\mathbb{R}} \varphi(z+t)^2 p_0(z)\, dz < \infty$ for $t \in \{-t_0, t_0\}$.

In particular, if L is a symmetric and twice differentiable convex function whose derivative is strictly increasing and bounded, then (B) holds and $\int_{\mathbb{R}} \varphi p_0 = 0$ for all symmetric densities p_0. Under the assumptions above, we first prove the $L^2(P_0)$-consistency of the initial estimates

$\tilde{\psi}_{n,j}$ of the score function ψ_0, from which it follows that the antitonic functions $\hat{\psi}_{n,j}$ consistently estimate the population-level projected score ψ_0^* in $L^2(P_0)$; see Lemmas S13 and S14 in Section S3.2. This enables us to establish the following result.

THEOREM 13. *Assume that* (A1)–(A5) *hold for the linear model* (29) *with symmetric error density* p_0. *If* $\bar{\beta}_n^{(j)} - \beta_0 = O_p(n^{-1/2})$ *for* $j \in \{1, 2, 3\}$, *then for any sequence of estimators* $(\hat{\beta}_n^{\mathrm{sym}})$ *with* $\hat{\beta}_n^{\mathrm{sym}} \in \{\hat{\beta}_n^{\dagger}, \hat{\beta}_n^{\ddagger}\}$ *for each* n, *we have*

$$\sqrt{n}(\hat{\beta}_n^{\mathrm{sym}} - \beta_0) \xrightarrow{d} N_d\biggl(0, \frac{\{\mathbb{E}(X_1 X_1^\top)\}^{-1}}{i^*(p_0)}\biggr)$$

as $n \to \infty$.

Theorem 13 reveals that our semiparametric convex M-estimators $\hat{\beta}_n^{\dagger}, \hat{\beta}_n^{\ddagger}$ are $\sqrt{n}$-consistent and have the same limiting Gaussian distribution as the "oracle" convex M-estimator $\hat{\beta}_{\psi_0^*} := \operatorname{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \ell_0^*(Y_i - X_i^\top \beta)$, where ℓ_0^* denotes an optimal convex loss function with right derivative ψ_0^*.

To understand the form of the asymptotic covariance matrix in Theorem 13, observe that since X_1 and ε_1 are independent, (X_1, Y_1) has joint density $(x, y) \mapsto p_0(y - x^\top \beta_0)$ with respect to the product measure $P_X \otimes \mathrm{Leb}$ on $\mathbb{R}^d \times \mathbb{R}$, where we write P_X for the distribution of X_1, and Leb for Lebesgue measure on $\mathbb{R}$. Therefore, in a parametric model where p_0 is known, the score function $\dot{\ell}_{\beta_0} \colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ for β_0 is given by

$$\dot{\ell}_{\beta_0}(x, y) := -x \psi_0(y - x^\top \beta_0),$$

where $\psi_0 = p_0'/p_0$. If $i(p_0) = \mathbb{E}(\psi_0(\varepsilon_1)^2)$ is finite, then because $\mathbb{E}\psi_0(\varepsilon_1) = 0$, the Fisher information matrix for β_0 is

$$I_{\beta_0} := \mathrm{Cov}\, \dot{\ell}_{\beta_0}(X_1, Y_1) = \mathrm{Cov}\{X_1 \psi_0(\varepsilon_1)\} = \mathbb{E}(X_1 X_1^\top) i(p_0) \in \mathbb{R}^{d \times d}. \tag{35}$$

Since $i(p_0) \geq i^*(p_0) > 0$ by Theorem 2*(d)*, I_{β_0} is positive definite if and only if $\mathbb{E}(X_1 X_1^\top)$ is positive definite. In this case, the convolution and local asymptotic minimax theorems (van der Vaart (1998), Chapter 8) indicate that $\sqrt{n}(\hat{\beta}^{\mathrm{MLE}} - \beta_0)$ in (4) has the "optimal" limiting distribution

$$N_d(0, I_{\beta_0}^{-1}) = N_d\biggl(0, \frac{\{\mathbb{E}(X_1 X_1^\top)\}^{-1}}{i(p_0)}\biggr)$$

among all (regular) sequences of estimators of β_0. By analogy with the previous display, the limiting covariance in Theorem 13 can be written as the inverse $(I_{\beta_0}^*)^{-1}$ of the *antitonic information matrix* $I_{\beta_0}^* := \mathbb{E}(X_1 X_1^\top) i^*(p_0)$. By Theorem 2, $1/i^*(p_0) = V_{p_0}(\psi_0^*) = \min_{\psi \in \Psi_\downarrow(p_0)} V_{p_0}(\psi)$, so by Proposition S24, $(I_{\beta_0}^*)^{-1}$ is the smallest possible limiting covariance among all convex M-estimators $\hat{\beta}_\psi$ based on a fixed $\psi \in \Psi_\downarrow(p_0)$. We can therefore interpret $(I_{\beta_0}^*)^{-1}$ as an *antitonic efficiency lower bound*.

3.4. *Linear regression with an intercept term.* For $d \geq 2$, now consider the linear model

$$Y_i = \mu_0 + \tilde{X}_i^\top \theta_0 + \varepsilon_i \quad \text{for } i \in [n], \tag{36}$$

where μ_0 is an explicit intercept term, so that $\beta_0 = (\theta_0, \mu_0)$ and $X_i = (\tilde{X}_i, 1)$ in (29) for $i \in [n]$. In the absence of further restrictions on the distribution of ε_1, the intercept term in this model is nonidentifiable since we may add a scalar to μ_0 and make a corresponding location shift to the distribution of ε_1 without changing the distribution of (X_1, Y_1). One could restore identifiability by including an assumption that $\mathbb{E}(\varepsilon_1) = 0$. However, to incorporate the

potential for heavy-tailed error distributions without a finite first moment (such as the Cauchy distribution), we instead impose a more general centring condition of the form $\mathbb{E}\zeta(\varepsilon_1) = 0$ for some prespecified decreasing function $\zeta : \mathbb{R} \to \mathbb{R}$ that satisfies condition (B) with $\varphi = \zeta$. In this case, define

$$\upsilon_{p_0} := V_{p_0}(\zeta) \in (0, \infty).$$

Naturally, taking ζ to be the function $z \mapsto -z$ yields the usual mean-zero assumption on the errors; on the other hand, for $\tau \in (0,1)$, letting $\zeta(z) = \mathbb{1}_{\{z<0\}} - \tau$ for $z \in \mathbb{R}$ constrains the errors to have τ-quantile equal to 0. In these two examples, $\upsilon_{p_0} = \mathbb{E}(\varepsilon_1^2)$ and $\upsilon_{p_0} = \tau(1-\tau)/p_0(0)^2$, respectively, and when $\upsilon_{p_0} \in (0, \infty)$, we automatically have $\int_{\mathbb{R}} \zeta(z+t)^2 p_0(z)\, dz < \infty$ for all $t \in \mathbb{R}$. In general, the fact that $\int_{\mathbb{R}} p_0\, d\zeta \in (-\infty, 0)$ ensures that $\mathbb{E}\zeta(\varepsilon_1 - c) = 0$ if and only if $c = 0$, and hence that μ_0 is identified by the equation $\mathbb{E}\zeta(Y_1 - \mu_0 - \tilde{X}_1^\top \theta_0) = 0$; see (S69).

Similar to Section 3.3, we employ three-fold cross-fitting with the convention $I_{j+3} = I_j$ for $j \in \{1,2\}$, and obtain pilot estimators $\bar{\beta}_n^{(j)} = (\bar{\theta}_n^{(j)}, \bar{\mu}_n^{(j)})$ of $\beta_0 = (\theta_0, \mu_0)$ given by (30) for $j \in \{1,2,3\}$ based on a fixed loss function L. We require the estimators $\bar{\theta}_n^{(j)}$ to be $\sqrt{n}$-consistent for θ_0 as $n \to \infty$. This is guaranteed by Proposition S24 if condition (B) is satisfied by $\varphi = -L^{(\mathrm{R})}$, so here we can either take L to be a twice differentiable convex loss function with a strictly decreasing and bounded derivative, or let L be a negative antiderivative of ζ.

We make some modifications to subsequent steps of the previous antitonic score matching procedure.

3′. *Antitonic projected score estimation*: For $j \in \{1,2,3\}$ and $i \in I_{j+1}$, use the out-of-sample residuals $\hat{\varepsilon}_i := Y_i - \tilde{X}_i^\top \bar{\theta}_n^{(j)}$ to construct the initial kernel-based score estimator $\tilde{\psi}_{n,j}$ and its antitonic projection $\hat{\psi}_{n,j} := \widehat{\mathcal{M}}_{\mathrm{R}}(\tilde{\psi}_{n,j} \circ \tilde{F}_{n,j}^{-1}) \circ \tilde{F}_{n,j}$ as before. Since p_0 is not symmetric in general, we use $\hat{\psi}_{n,j}$ instead of $\hat{\psi}_{n,j}^{\mathrm{anti}}$.

4′. *Plug-in cross-fitted convex M-estimator*: For $j \in \{1,2,3\}$, let $\hat{\ell}_{n,j} : \mathbb{R} \to \mathbb{R}$ be the induced convex loss function given by $\hat{\ell}_{n,j}(z) := -\int_0^z \hat{\psi}_{n,j}$, and define

$$\hat{\theta}_n^{(j)} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{d-1}} \sum_{i \in I_{j+2}} \hat{\ell}_{n,j}\big(Y_i - \bar{X}_{n,j}^\top \bar{\theta}_n^{(j)} - (\tilde{X}_i - \bar{X}_{n,j})^\top \theta\big), \tag{37}$$

where $\bar{X}_{n,j} := |I_{j+2}|^{-1} \sum_{i \in I_{j+2}} \tilde{X}_i$. Finally, let $\hat{\theta}_n^\dagger := (\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)} + \hat{\theta}_n^{(3)})/3$. Alternatively, define

$$\hat{\theta}_n^\ddagger \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{d-1}} \sum_{j=1}^{3} \sum_{i \in I_{j+2}} \hat{\ell}_{n,j}\big(Y_i - \bar{X}_{n,j}^\top \bar{\theta}_n^{(j)} - (\tilde{X}_i - \bar{X}_{n,j})^\top \theta\big).$$

5′. *Intercept estimation*: Taking either $\hat{\theta}_n = \hat{\theta}_n^\dagger$ or $\hat{\theta}_n = \hat{\theta}_n^\ddagger$, let

$$\hat{\mu}_n^\zeta \in \operatorname*{argmin}_{\mu \in \mathbb{R}} \sum_{i=1}^{n} L_\zeta\big(Y_i - \tilde{X}_i^\top \hat{\theta}_n - \mu\big), \tag{38}$$

where L_ζ is a negative antiderivative of ζ. Finally, output $\hat{\beta}_n^\zeta := (\hat{\theta}_n, \hat{\mu}_n^\zeta)$.

In summary, $\hat{\theta}_n$ minimises an empirical risk based on centred covariates and an estimated convex loss (via antitonic score matching), while $\hat{\mu}_n^\zeta$ is defined as a location M-estimator with respect to the residuals from $\hat{\theta}_n$ and the fixed loss function used to centre the regression errors. There exists a minimiser $\hat{\theta}_n^{(j)}$ in (37) if $\inf_{z \in \mathbb{R}} \hat{\psi}_{n,j}(z) < 0 < \sup_{z \in \mathbb{R}} \hat{\psi}_{n,j}(z)$, and $\hat{\theta}_n^{(j)}$ is unique if $\hat{\psi}_{n,j}$ is strictly decreasing and the design matrix has full column rank; see Proposition S24*(a)*.

Step 4′ can be motivated by semiparametric calculations that are summarised briefly below and presented formally in Section S3.5. We verify in Proposition S22*(a)* that if μ_0 is viewed as a nuisance parameter, then the *efficient score function* $\tilde{\ell}_{\theta_0}: \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d-1}$ (van der Vaart (1998), Section 25.4) for θ_0 is given by

$$\tilde{\ell}_{\theta_0}(x, y) := -(\tilde{x} - \tilde{m})\psi_0\big(y - x^\top \beta_0\big), \tag{39}$$

where $\tilde{m} := \mathbb{E}(\tilde{X}_1) \in \mathbb{R}^{d-1}$, $x = (x_1, \ldots, x_d)$ and $\tilde{x} = (x_1, \ldots, x_{d-1})$. This is the same regardless of whether p_0 is known or unknown. The *efficient information matrix* for θ_0 is

$$\tilde{I}_{\theta_0} := \mathbb{E}\big(\tilde{\ell}_{\theta_0}(X_1, Y_1)\tilde{\ell}_{\theta_0}(X_1, Y_1)^\top\big) = \mathbb{E}\big(\psi_0(\varepsilon_1)^2\big)\,\mathrm{Cov}(\tilde{X}_1) = i(p_0)\Sigma, \tag{40}$$

where $\Sigma := \mathrm{Cov}(\tilde{X}_1)$, and the semiparametric efficiency lower bound (e.g., van der Vaart (1998), p. 367) is $\tilde{I}_{\theta_0}^{-1} = \Sigma^{-1}/i(p_0)$. This is the top-left $(d-1) \times (d-1)$ submatrix of $I_{\beta_0}^{-1}$ in (35), and is asymptotically attained by an adaptive estimator $\tilde{\theta}_n$ of θ' that solves a version of the *efficient score equations* $\sum_{i=1}^n \tilde{\ell}_{\tilde{\theta}_n}(X_i, Y_i) = 0$ with ψ_0 replaced with an $L^2(P_0)$-consistent score estimate; see Bickel (1982), Example 3 and van der Vaart (1998), Chapter 25.8. In (37), we instead target a surrogate of the efficient score $\tilde{\ell}_{\theta_0}$ above with ψ_0 replaced by ψ_0^*. If $\hat{\psi}_{n,j}$ is continuous, then any minimiser $\hat{\theta}_n^{(j)}$ satisfies

$$\sum_{i \in I_{j+2}} (\tilde{X}_i - \bar{X}_{n,j}) \cdot \hat{\psi}_{n,j}\big(Y_i - \bar{X}_{n,j}^\top \bar{\theta}_n^{(j)} - (\tilde{X}_i - \bar{X}_{n,j})^\top \hat{\theta}_n^{(j)}\big) = 0,$$

which is a variant of the efficient score equations for θ_0. Since we do not assume knowledge of the population mean $\tilde{m} = \mathbb{E}(\tilde{X}_1)$, we replace it with its sample analogue $\bar{X}_{n,j}$ in each of the three folds. The rationale for the final Step 5′ is that $\hat{\beta}_n^\zeta = (\hat{\theta}_n, \hat{\mu}_n^\zeta)$ solves a version of the efficient score equations for β_0 when p_0 is regarded as a nuisance parameter; see (S84) in Section S3.5.

Under the regularity conditions in Section 3.3, it turns out that, up to a translation by μ_0, the projected score functions $\hat{\psi}_{n,j}$ are also $L^2(P_0)$-consistent estimators of ψ_0^* in this setting (Lemma S14). It follows that with probability tending to 1 as $n \to \infty$, we have $\lim_{z \to -\infty} \hat{\psi}_{n,j}(z) > 0 > \lim_{z \to \infty} \hat{\psi}_{n,j}(z)$, and hence $\hat{\theta}_n^\dagger$ and $\hat{\theta}_n^\ddagger$ exist. We can then adapt the proof strategy for Theorem 13 above to establish the following main result.

THEOREM 14. *Assume that* (A1)–(A5) *hold and, moreover, that* (B) *is satisfied by* $\varphi = \zeta$ *in the linear model* (36) *with* $\mathbb{E}\zeta(\varepsilon_1) = 0$. *Suppose further that* $\bar{\theta}_n^{(j)} - \theta_0 = O_p(n^{-1/2})$ *for* $j \in \{1, 2, 3\}$. *Then for any sequence of estimators* $\hat{\beta}_n^\zeta = (\hat{\theta}_n, \hat{\mu}_n^\zeta)$ *such that* $\hat{\theta}_n \in \{\hat{\theta}_n^\dagger, \hat{\theta}_n^\ddagger\}$ *for each* n, *we have*

$$\sqrt{n}(\hat{\beta}_n^\zeta - \beta_0) \xrightarrow{d} N_d\big(0, (\tilde{I}_{\beta_0}^*)^{-1}\big);$$

here,

$$(\tilde{I}_{\beta_0}^*)^{-1} = \begin{pmatrix} \dfrac{\Sigma^{-1}}{i^*(p_0)} & -\dfrac{\Sigma^{-1}\tilde{m}}{i^*(p_0)} \\ -\dfrac{\tilde{m}^\top \Sigma^{-1}}{i^*(p_0)} & \upsilon_{p_0} + \dfrac{\tilde{m}^\top \Sigma^{-1}\tilde{m}}{i^*(p_0)} \end{pmatrix} \in \mathbb{R}^{d \times d} \tag{41}$$

is the inverse of

$$\begin{aligned} \tilde{I}_{\beta_0}^* &:= \begin{pmatrix} i^*(p_0)\Sigma + \tilde{m}\tilde{m}^\top/\upsilon_{p_0} & \tilde{m}/\upsilon_{p_0} \\ \tilde{m}^\top/\upsilon_{p_0} & 1/\upsilon_{p_0} \end{pmatrix} \\ &= i^*(p_0)\mathbb{E}\big(X_1 X_1^\top\big) - \bigg(i^*(p_0) - \frac{1}{\upsilon_{p_0}}\bigg)\mathbb{E}(X_1)\mathbb{E}(X_1)^\top. \end{aligned}$$

As an immediate consequence of Theorem 14,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_{d-1}\Big(0, \frac{\Sigma^{-1}}{i^*(p_0)}\Big) \tag{42}$$

as $n \to \infty$, and the proof reveals that this holds even without the centring constraint $\mathbb{E}\zeta(\varepsilon_1) = 0$; see also Proposition S22*(a)*. Therefore, $\hat{\theta}_n$ is an data-driven convex M-estimator of θ_0 whose limiting covariance $\Sigma^{-1}/i^*(p_0)$ is identical to $\tilde{I}_{\theta_0}^{-1}$ in (40), except with $i(p_0)$ in place of $i^*(p_0)$. Similarly, $\tilde{I}^*_{\beta_0}$ is the antitonic analogue of the efficient information matrix $\tilde{I}_{\beta_0}$ for β_0; see (S81) in Proposition S22*(b)*. Consequently, we can view Theorem 14 as an "antitonic efficiency" result.

To interpret the bottom-right entry in (41), first suppose that θ_0 is known and p_0 is unknown. Then the problem of estimating μ_0 reduces to one-dimension location estimation based on $\tilde{Y}_i := Y_i - \tilde{X}_i^\top \theta_0 = \mu_0 + \varepsilon_i$ for $i \in [n]$. Since $\mathbb{E}\zeta(\tilde{Y}_1 - \mu_0) = 0$, the corresponding semiparametric efficiency lower bound is υ_{p_0}, which is the one-dimensional version of (3); see van der Vaart (1998), Example 25.24, for the special case of mean estimation. On the other hand, when θ_0 is also unknown, the efficiency lower bound for μ_0 is instead the bottom-right entry $(\tilde{I}_{\beta_0}^{-1})_{d,d} = \upsilon_{p_0} + \tilde{m}^\top \Sigma^{-1} \tilde{m}/i(p_0)$ in (S81). Therefore, it is strictly harder to estimate μ_0 in our semiparametric setting unless $\tilde{m} = 0$, and the asymptotic variance of our $\hat{\mu}_n^\zeta$ is the antitonic counterpart of $(\tilde{I}_{\beta_0}^{-1})_{d,d}$. Observe also that $i^*(p_0) \geq 1/\upsilon_{p_0}$ by (16), so the limiting covariance matrices in Theorems 13 and 14 satisfy $(I^*_{\beta_0})^{-1} \preceq (\tilde{I}^*_{\beta_0})^{-1}$ in the positive semidefinite (Loewner) ordering $\preceq$, with equality if and only if ζ is proportional to $-\psi_0^*$.

Lemma S23 in Section S3.6 establishes that $\hat{\theta}_n$ always has asymptotic efficiency at least that of the composite quantile estimator (Zou and Yuan (2008)), and also that there exist log-concave densities p_0 for which the latter has arbitrarily low efficiency relative to the former.

3.5. *Inference.* To perform asymptotically valid inference for β_0 based on Theorems 13 and 14 when p_0 is unknown, we require a consistent estimator of the antitonic information $i^*(p_0)$. This can be constructed using residuals $\check{\varepsilon}_i := Y_i - X_i^\top \bar{\beta}_n^{(j)}$ for $j \in \{1, 2, 3\}$ and $i \in I_{j+2}$ in place of the unobserved errors, where $\bar{\beta}_n^{(j)}$ are $\sqrt{n}$-consistent estimators of β_0 given by (30).

LEMMA 15. *In the setting of Theorem 13, let* $\check{\psi}_{n,j} \in \{\hat{\psi}_{n,j}, \hat{\psi}^{\mathrm{anti}}_{n,j}\}$ *for* $n \in \mathbb{N}$ *and* $j \in \{1, 2, 3\}$. *Then*

$$\hat{\imath}_n := \frac{1}{n} \sum_{j=1}^{3} \sum_{i \in I_{j+2}} \check{\psi}_{n,j}(\check{\varepsilon}_i)^2 \xrightarrow{p} i^*(p_0).$$

The same conclusion holds in the setting of Theorem 14 if instead $\check{\psi}_{n,j}(\cdot) = \hat{\psi}_{n,j}(\cdot + \bar{\mu}_n^{(j)})$ *for all* n, j.

Therefore, $I^*_{\beta_0} = \mathbb{E}(X_1 X_1^\top) i^*(p_0)$ can be estimated consistently by

$$\hat{I}_n := \frac{\hat{\imath}_n}{n} \sum_{i=1}^{n} X_i X_i^\top = \frac{\hat{\imath}_n}{n} X^\top X,$$

where $X = (X_1 \cdots X_n)^\top \in \mathbb{R}^{n \times d}$ has full column rank with probability tending to 1 under (A5). When p_0 is symmetric, it follows from Theorem 13 and Lemma 15 that

$$\sqrt{n} \hat{I}_n^{1/2} (\hat{\beta}_n^{\mathrm{sym}} - \beta_0) \xrightarrow{d} N_d(0, I_d).$$

Thus, writing $\hat{\beta}^{\mathrm{sym}}_{n,j}$ for the jth component of $\hat{\beta}^{\mathrm{sym}}_n$ and $\hat{s}_j := (\hat{I}_n^{-1})_{jj}$ for the jth diagonal entry of $\hat{I}_n^{-1}$, we deduce that

$$\left[\hat{\beta}^{\mathrm{sym}}_{n,j} - z_{\alpha/2}\sqrt{\frac{\hat{s}_j}{n}}, \hat{\beta}^{\mathrm{sym}}_{n,j} + z_{\alpha/2}\sqrt{\frac{\hat{s}_j}{n}}\right]$$

is an asymptotic $(1-\alpha)$-level confidence interval for the jth component of β_0, where $z_{\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution. Moreover,

$$\{b \in \mathbb{R}^d : n(\hat{\beta}^{\mathrm{sym}}_n - b)^\top \hat{I}_n(\hat{\beta}^{\mathrm{sym}}_n - b) \le \chi^2_d(\alpha)\}$$

is an asymptotic $(1-\alpha)$-confidence ellipsoid for β_0, where $\chi^2_d(\alpha)$ denotes the $(1-\alpha)$-quantile of the χ^2_d distribution.

On the other hand, in the setting of Section 3.4, $\tilde{I}_n := n^{-1}\sum_{i=1}^n \hat{\imath}_n(\tilde{X}_i - \bar{\tilde{X}}_n)(\tilde{X}_i - \bar{\tilde{X}}_n)^\top$ is a consistent estimator of $i^*(p_0)\Sigma$, where $\bar{\tilde{X}}_n := n^{-1}\sum_{i=1}^n \tilde{X}_i$. In view of (42), we can therefore construct asymptotically valid confidence sets for θ_0 and confidence intervals for its components, as above. To obtain a sample approximation to the asymptotic variance of the intercept estimate $\hat{\mu}^\zeta_n$, we also require a consistent estimate $\hat{\upsilon}_n$ of $\upsilon_{p_0} = V_{p_0}(\zeta)$. If the errors are mean-centred via $\zeta : z \mapsto -z$, then a natural estimator of $\upsilon_{p_0} = \mathbb{E}(\varepsilon_1^2)$ is $n^{-1}\sum_{i=1}^n \check{\varepsilon}_i^2$. On the other hand, if the τ-quantile of the errors are centred via $\zeta : z \mapsto \mathbb{1}_{\{z<0\}} - \tau$ for some $\tau \in (0,1)$, then we can approximate $\upsilon_{p_0} = \tau(1-\tau)/p_0(0)^2$ using a kernel density estimator of $p_0(0)$. The following lemma verifies that using the residuals $\check{\varepsilon}_1, \ldots, \check{\varepsilon}_n$ is justified more generally in this context.

LEMMA 16. *Suppose that $\zeta = \zeta_{\mathrm{ac}} - \sum_{m=1}^M \zeta_m \mathbb{1}_{[z_m,\infty)}$ for some $M \in \mathbb{N}_0$, where ζ_{ac} is absolutely continuous on $\mathbb{R}$, while $\zeta_m > 0$ and $z_m \in \mathbb{R}$ for all $m \in [M]$. Define $\tilde{p}_n : \mathbb{R} \to \mathbb{R}$ by $\tilde{p}_n(z) := n^{-1}\sum_{i=1}^n K_h(z - \check{\varepsilon}_i)$ for some square-integrable kernel K and bandwidth $h \equiv h_n$ satisfying $h_n \to 0$ and $nh_n \to \infty$. If ζ'_{ac} is continuous Lebesgue almost everywhere on $\mathbb{R}$, then under the hypotheses of Theorem 14,*

$$\hat{\upsilon}_n := \frac{n^{-1}\sum_{i=1}^n \zeta(\check{\varepsilon}_i)^2}{\{n^{-1}\sum_{i=1}^n \zeta'_{\mathrm{ac}}(\check{\varepsilon}_i) - \sum_{m=1}^M \zeta_m \tilde{p}_n(z_m)\}^2} \xrightarrow{p} \upsilon_{p_0}.$$

Now define

$$\hat{I}^\zeta_n := \hat{I}_n - \left(\hat{\imath}_n - \frac{1}{\hat{\upsilon}_n}\right)\bar{X}_n\bar{X}_n^\top,$$

where $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$, and denote by $\tilde{s}_j$ the jth diagonal entry of $(\hat{I}^\zeta_n)^{-1}$ for $j \in [d]$. We deduce from Theorem 14 together with Lemmas 15 and 16 that

$$\left[\hat{\beta}^\zeta_{n,j} - z_{\alpha/2}\sqrt{\frac{\tilde{s}_j}{n}}, \hat{\beta}^\zeta_{n,j} + z_{\alpha/2}\sqrt{\frac{\tilde{s}_j}{n}}\right]$$

is an asymptotic $(1-\alpha)$-level confidence interval for the jth component of β_0, and

$$\{b \in \mathbb{R}^d : n(\hat{\beta}^\zeta_n - b)^\top \hat{I}^\zeta_n(\hat{\beta}^\zeta_n - b) \le \chi^2_d(\alpha)\} \tag{43}$$

is an asymptotic $(1-\alpha)$-confidence ellipsoid for β_0. Lemma 15 also ensures that standard linear model diagnostics, either based on heuristics such as Cook's distance (Cook (1977)), or formal goodness-of-fit tests (Janková et al. (2020)), can be applied.

4. Numerical experiments. In our numerical experiments, we generate covariates $\tilde{X}_1, \ldots, \tilde{X}_n \overset{\text{iid}}{\sim} N_{d-1}(\mathbf{1}_{d-1}, I_{d-1})$, where $\mathbf{1}_{d-1}$ denotes the $(d-1)$-dimensional all-ones vector, and responses $Y_1, \ldots, Y_n$ according to the linear model (36). Independently of the covariates, we draw errors $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} P_0$ for the following choices of P_0:

(i) P_0 is standard Gaussian.
(ii) P_0 is standard Cauchy.
(iii) Gaussian scale mixture: $P_0 = \frac{1}{2}N(0,1) + \frac{1}{2}N(0,16)$.
(iv) Gaussian location mixture: $P_0 = \frac{1}{2}N(-\frac{3}{2}, \frac{1}{100}) + \frac{1}{2}N(\frac{3}{2}, \frac{1}{100})$. This distribution is similar to the one constructed in the proof of Proposition 7 to show that log-concave maximum likelihood estimation of the error distribution may result in arbitrarily large efficiency loss.
(v) Smoothed uniform: P_0 is the distribution of $U + \frac{1}{10}Z$, where $U \sim \text{Unif}[-1,1]$ and $Z \sim N(0,1)$ are independent.
(vi) Smoothed exponential: P_0 is the distribution of $W - 1 + \frac{\sqrt{3}}{10}Z$, where $W \sim \text{Exp}(1)$ and $Z \sim N(0,1)$ are independent. We choose the standard deviation $\frac{\sqrt{3}}{10}$ for the Gaussian component so that the ratio of the variances of the non-Gaussian and the Gaussian components is the same as that in the smoothed uniform setting.

In cases (i), (v) and (vi), P_0 is log-concave, while for the other settings it is not. Since the OLS and LAD estimators are targeting different population intercepts in cases where the mean and median of P_0 are not equal (so at most one of these estimators can be Fisher consistent), we focus on estimation of the $(d-1)$-dimensional subvector θ_0 of β_0 in order to present a fair comparison.

We compare the performance of two versions of our procedure with an oracle approach and four existing methods. The first variant of our procedure, which we refer to as **ASM** (antitonic score matching) in all of the plots, is as described in Section 3.4, except that we do not perform sample splitting, cross-fitting or truncation of the initial score estimates. These devices are convenient for theoretical analysis but not essential in practice. More precisely, **ASM** first constructs a pilot estimator $\bar{\beta}_n = (\bar{\theta}_n, \bar{\mu}_n)$, and then uses the vector of residuals $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$, where $\hat{\varepsilon}_i := Y_i - \tilde{X}_i^\top \bar{\theta}_n$, to obtain an initial kernel-based score estimator $\tilde{\psi}_n$, formed using a Gaussian kernel and the default Silverman's choice of bandwidth (Silverman (1986), p. 48). Following Step 3′ in Section 3.4, we estimate the antitonic projected score and the corresponding convex loss function by $\hat{\psi}_n := \widehat{\mathcal{M}}_{\mathrm{R}}(\tilde{\psi}_n \circ \tilde{F}_n^{-1}) \circ \tilde{F}_n$ and $\hat{\ell}_n$, respectively, where $\tilde{F}_n$ denotes the distribution function associated with the kernel density estimate. Finally, we use Newton's algorithm with Hessian modification[6] (Nocedal and Wright (2006), Section 3.4) to compute our semiparametric estimator

$$\hat{\theta}_n^{\text{ASM}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{d-1}} \sum_{i=1}^n \hat{\ell}_n(Y_i - \bar{\tilde{X}}_n^\top \bar{\theta}_n - \tilde{W}_i^\top \theta),$$

where $\bar{\tilde{X}}_n := n^{-1}\sum_{i=1}^n \tilde{X}_i$ and $\tilde{W}_i := \tilde{X}_i - \bar{\tilde{X}}_n$ for $i \in [n]$.

In the second version of our procedure, which we refer to as **Alt** in our plots, we implement an empirical analogue of the alternating optimisation procedure described in Section 3.2. We start with an uninformative initialiser $(\hat{\theta}_n^{(0)}, \hat{\mu}_n^{(0)}) = (0,0) \in \mathbb{R}^{d-1} \times \mathbb{R}$, and then alternate between the following steps for $t \in \mathbb{N}$:

I. Compute residuals $\hat{\varepsilon}_i^{(t-1)} := Y_i - \hat{\mu}_n^{(t-1)} - \tilde{X}_i^\top \hat{\theta}_n^{(t-1)}$ for $i \in [n]$, and hence estimate the antitonic projected score $\hat{\psi}_n^{(t-1)}$ and the corresponding convex loss $\hat{\ell}_n^{(t-1)}$ as for **ASM**.

[6]This modification involves adding the identity matrix to the Hessian prior to its version.

II. Update $(\hat{\theta}_n^{(t)}, \hat{\mu}_n^{(t)}) \in \operatorname{argmin}_{(\theta,\mu)\in\mathbb{R}^{d-1}\times\mathbb{R}} \sum_{i=1}^n \hat{\ell}_n^{(t-1)}(Y_i - \mu - \tilde{X}_i^\top\theta)$.

We iterate these steps until convergence of the empirical score matching objective

$$\hat{D}_n(\hat{\psi}_n^{(t)}; \hat{\varepsilon}_1^{(t)}, \ldots, \hat{\varepsilon}_n^{(t)}) = \frac{1}{n}\sum_{i=1}^n \{\hat{\psi}_n^{(t)}(\hat{\varepsilon}_i^{(t)}) + 2(\hat{\psi}_n^{(t)})'(\hat{\varepsilon}_i^{(t)})\}$$

defined in (26); in all of our experiments both $\hat{\theta}_n^{(t)}$ and the score matching objective values did indeed converge.

The alternative approaches that we consider are as follows:

- **Oracle**: The M-estimator $\hat{\theta}_n^{\mathrm{oracle}}$, where

$$\begin{pmatrix}\hat{\theta}_n^{\mathrm{oracle}}\\ \hat{\mu}_n^{\mathrm{oracle}}\end{pmatrix} \in \operatorname*{argmin}_{(\theta,\mu)\in\mathbb{R}^{d-1}\times\mathbb{R}} \sum_{i=1}^n \ell_0^*(Y_i - \mu - \tilde{X}_i^\top\theta)$$

  is defined with respect to the optimal convex loss function ℓ_0^*. Although $\hat{\theta}_n^{\mathrm{oracle}}$ is antitonically efficient in the sense of (42), it is not a valid estimator in our semiparametric framework since it requires knowledge of p_0.
- **LAD**: The least absolute deviation estimator $\hat{\theta}_n^{\mathrm{LAD}}$, where

$$\hat{\beta}_n^{\mathrm{LAD}} = \begin{pmatrix}\hat{\theta}_n^{\mathrm{LAD}}\\ \hat{\mu}_n^{\mathrm{LAD}}\end{pmatrix} \in \operatorname*{argmin}_{(\theta,\mu)\in\mathbb{R}^{d-1}\times\mathbb{R}} \sum_{i=1}^n |Y_i - \mu - \tilde{X}_i^\top\theta|.$$

- **OLS**: The ordinary least squares estimator $\hat{\theta}_n^{\mathrm{OLS}}$, where

$$\hat{\beta}_n^{\mathrm{OLS}} = \begin{pmatrix}\hat{\theta}_n^{\mathrm{OLS}}\\ \hat{\mu}_n^{\mathrm{OLS}}\end{pmatrix} \in \operatorname*{argmin}_{(\theta,\mu)\in\mathbb{R}^{d-1}\times\mathbb{R}} \sum_{i=1}^n (Y_i - \mu - \tilde{X}_i^\top\theta)^2.$$

- **1S**: The semiparametric one-step method, where we start with a pilot estimator $(\bar{\theta}_n, \bar{\mu}_n)$, compute a (not necessarily decreasing) nonparametric score estimate, and then update $\bar{\theta}_n$ with a single Newton step instead of solving the estimating equations exactly. Our implementation follows van der Vaart (1998), Chapter 25.8: we split the data into two folds of equal size indexed by I_1 and I_2, and then use the residuals $\hat{\varepsilon}_i := Y_i - \tilde{X}_i^\top\bar{\theta}_n - \bar{\mu}_n$ for $i \in I_1$ and $i \in I_2$ separately to obtain kernel density estimates $\hat{p}_{n,1}$, $\hat{p}_{n,2}$ of p_0 (constructed as for **ASM**). Defining the score estimates $\hat{\psi}_{n,j} := \hat{p}'_{n,j}/\hat{p}_{n,j}$ for $j \in \{1, 2\}$, we output the cross-fitted estimator $\hat{\theta}_n^{\mathrm{1S}}$, where

$$\hat{\theta}_n^{\mathrm{1S}} := \bar{\theta}_n - \Bigg(\sum_{i\in I_1} \hat{\psi}_{n,2}(\hat{\varepsilon}_i)^2 \tilde{W}_i \tilde{W}_i^\top + \sum_{i\in I_2} \hat{\psi}_{n,1}(\hat{\varepsilon}_i)^2 \tilde{W}_i \tilde{W}_i^\top\Bigg)^{-1}$$
$$\times \Bigg(\sum_{i\in I_1} \hat{\psi}_{n,2}(\hat{\varepsilon}_i)\tilde{W}_i + \sum_{i\in I_2} \hat{\psi}_{n,1}(\hat{\varepsilon}_i)\tilde{W}_i\Bigg).$$

- **LCMLE**: We estimate the error density p_0 using the log-concave maximum likelihood estimator (Cule, Samworth and Stewart (2010), Dümbgen, Samworth and Schuhmacher (2011, 2013)). More precisely, again writing $\mathcal{P}_{\mathrm{LC}}$ for the set of univariate log-concave densities and defining $Q(\theta, \mu; p) := \sum_{i=1}^n \log p(Y_i - \mu - \tilde{X}_i^\top\theta)$ for $\theta \in \mathbb{R}^{d-1}$, $\mu \in \mathbb{R}$ and $p \in \mathcal{P}_{\mathrm{LC}}$, we start with a pilot estimator $(\hat{\theta}_n^{(0)}, \hat{\mu}_n^{(0)})$ and alternate the following two steps for $t \in \mathbb{N}$ until convergence of $Q(\hat{\theta}_n^{(t)}, \hat{\mu}_n^{(t)}; \hat{p}_n^{(t)})$:

$$\hat{p}_n^{(t)} \in \operatorname*{argmax}_{p\in\mathcal{P}_{\mathrm{LC}}} Q(\hat{\theta}_n^{(t-1)}, \hat{\mu}_n^{(t-1)}; p), \qquad \begin{pmatrix}\hat{\theta}_n^{(t)}\\ \hat{\mu}_n^{(t)}\end{pmatrix} \in \operatorname*{argmax}_{(\theta,\mu)\in\mathbb{R}^{d-1}\times\mathbb{R}} Q(\theta, \mu; \hat{p}_n^{(t)}).$$

TABLE 1
Squared estimation error ($\times 10^3$) *for different estimators of* θ_0, *with* $n = 600$ *and* $d = 6$

	Oracle	ASM	Alt	LCMLE	1S	LAD	OLS
Standard Gaussian	8.51	8.96	9.01	9.43	9.77	12.70	**8.51**
Standard Cauchy	19.98	**20.44**	20.68	48.31	21.74	21.15	3.9×10^6
Gaussian scale mixture	30.71	**31.58**	31.78	34.49	36.48	34.67	73.45
Gaussian location mixture	0.17	0.18	**0.17**	0.74	18.28	332.08	18.72
Smoothed uniform	1.03	1.36	**1.18**	1.31	2.10	8.38	2.99
Smoothed exponential	1.91	2.24	**2.10**	2.26	3.33	8.60	8.86

For all of our error densities p_0, condition (B) is satisfied by $\varphi = -\operatorname{sgn}$, and hence by Proposition S24, $\hat{\beta}_n^{\mathrm{LAD}}$ is $\sqrt{n}$-consistent with asymptotic variance factor $V_{p_0}(\varphi) = 1/(4p_0(0)^2)$. Therefore, we took $\hat{\beta}_n^{\mathrm{LAD}}$ to be our pilot estimator for all methods except in the Gaussian location mixture setting (iv), where instead we used $\hat{\beta}_n^{\mathrm{OLS}}$ since $p_0(0)$ is close to 0, and hence $V_{p_0}(\varphi)$ for $\hat{\beta}_n^{\mathrm{LAD}}$ is very large.

We set $d = 6$, $\mu_0 = 2$ and drew θ_0 uniformly at random from the centred Euclidean sphere in $\mathbb{R}^{d-1}$ of radius 3. For each of the estimators above, we computed the average squared Euclidean norm errors $\|\hat{\theta}_n - \theta_0\|^2$ over 200 repetitions. We considered each error distribution in turn and compare the average squared estimation error when $n = 600$. The results are presented in Table 1. We observe that our proposed procedures ASM and Alt have the lowest estimation error except in the case of standard Gaussian error, where OLS coincides with the oracle convex loss estimator and has a slightly lower estimation error. It is interesting that the one-step estimator (1S) can perform very poorly in finite samples, and in particular, may barely improve on its initialiser. In all settings considered, ASM and Alt have comparable error. Further numerical results are presented in Section S6.

5. Discussion. One of the messages of this paper is that, despite the Gauss–Markov theorem, the success of ordinary least squares is relatively closely tied to Gaussian or near-Gaussian error distributions. Our antitonic score matching approach aims to free the practitioner from the Gaussian straitjacket while retaining the convenience and stability of working with convex loss functions. The Fisher divergence projection framework brings together previously disparate ideas on shape-constrained estimation, score matching, information theory and classical robust statistics. Given the prevalence of procedures in statistics and machine learning that are constructed as optimisers of prespecified loss functions, we look forward to seeing how related insights may lead to more flexible, data-driven and computationally feasible approaches that combine robustness and efficiency.

Acknowledgments. The authors thank David Firth, Elliot Young and Cun-Hui Zhang for helpful discussions, as well as an Associate Editor and two anonymous reviewers whose constructive comments led to several improvements in the paper.

Funding. The research of YCK and MX was supported by National Science Foundation Grants DMS-2311299 and DMS-2113671.

OYF and RJS were supported by Engineering and Physical Sciences Research Council Programme Grant EP/N031938/1, while RJS was also supported by European Research Council Advanced Grant 101019498.

SUPPLEMENTARY MATERIAL

Supplementary material for "Optimal convex M-estimation via score matching" (DOI: 10.1214/25-AOS2572SUPP; .pdf). The supplementary material contains the proofs of all theoretical results as well as additional simulations.

REFERENCES

AMARI, S. and NAGAOKA, H. (2000). *Methods of Information Geometry*. *Translations of Mathematical Monographs* **191**. Amer. Math. Soc., Providence, RI. MR1800071 https://doi.org/10.1090/mmono/191

ARCONES, M. A. (1998). Asymptotic theory for M-estimators over a convex kernel. *Econometric Theory* **14** 387–422. MR1650029 https://doi.org/10.1017/S0266466698144018

BARBER, R. F. and SAMWORTH, R. J. (2021). Local continuity of log-concave projection, with applications to estimation under model misspecification. *Bernoulli* **27** 2437–2472. MR4303890 https://doi.org/10.3150/20-BEJ1316

BARRON, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4331–4339.

BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110** 14563–14568.

BENTON, J., SHI, Y., DE BORTOLI, V., DELIGIANNIDIS, G. and DOUCET, A. (2024). From denoising diffusions to denoising Markov models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 286–301. MR4896938 https://doi.org/10.1093/jrsssb/qkae005

BERAN, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313. MR0518885

BETANCOURT, M., BYRNE, S., LIVINGSTONE, S. and GIROLAMI, M. (2017). The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli* **23** 2257–2298. MR3648031 https://doi.org/10.3150/16-BEJ810

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434. MR0386168

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671. MR0663424

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

BRUNEL, V.-E. (2023). Geodesically convex M-estimation in metric spaces. In *The Thirty Sixth Annual Conference on Learning Theory* 2188–2210. PMLR.

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

CELENTANO, M. and MONTANARI, A. (2022). Fundamental barriers to high-dimensional regression with convex penalties. *Ann. Statist.* **50** 170–196. MR4382013 https://doi.org/10.1214/21-aos2100

CHENG, X., CHATTERJI, N. S., BARTLETT, P. L. and JORDAN, M. I. (2018). Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory* 300–323. PMLR.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CHINOT, G., LECUÉ, G. and LERASLE, M. (2020). Robust statistical learning with Lipschitz and convex loss functions. *Probab. Theory Related Fields* **176** 897–940. MR4087486 https://doi.org/10.1007/s00440-019-00931-3

COOK, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* **19** 15–18. MR0436478 https://doi.org/10.2307/1268249

COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley, Hoboken, NJ. MR2239987

COX, D. D. (1985). A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Ann. Inst. Statist. Math.* **37** 271–288. MR0799240 https://doi.org/10.1007/BF02481097

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. MR2758237 https://doi.org/10.1111/j.1467-9868.2010.00753.x

DALALYAN, A. S., GOLUBEV, G. K. and TSYBAKOV, A. B. (2006). Penalized maximum likelihood and semiparametric second-order efficiency. *Ann. Statist.* **34** 169–201. MR2275239 https://doi.org/10.1214/009053605000000895

DERENSKI, J., FAN, Y., JAMES, G. and XU, M. (2023). An empirical Bayes shrinkage method for functional data. Submitted.

DE BORTOLI, V., MATHIEU, E., HUTCHINSON, M., THORNTON, J., TEH, Y. W. and DOUCET, A. (2022). Riemannian score-based generative modelling. *Adv. Neural Inf. Process. Syst.* **35** 2406–2422.

DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z

DONOHO, D. L. and MONTANARI, A. (2015). Variance breakdown of Huber M-estimators: $n/p \in (1, \infty)$. arXiv Preprint. Available at arXiv:1503.02106.

DOSS, C. R. and WELLNER, J. A. (2019). Univariate log-concave density estimation with symmetry or modal constraints. *Electron. J. Stat.* **13** 2391–2461. MR3983344 https://doi.org/10.1214/19-EJS1574

DÜMBGEN, L., SAMWORTH, R. and SCHUHMACHER, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730. MR2816336 https://doi.org/10.1214/10-AOS853

DÜMBGEN, L., SAMWORTH, R. J. and SCHUHMACHER, D. (2013). Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back: High-Dimensional Models and Processes*. *Inst. Math. Stat. (IMS) Collect.* **9** 78–90. IMS, Beachwood, OH. MR3186750 https://doi.org/10.1214/12-IMSCOLL907

EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860 https://doi.org/10.1198/jasa.2011.tm11181

EGGERMONT, P. P. B. and LARICCIA, V. N. (2000). Maximum likelihood estimation of smooth monotone and unimodal densities. *Ann. Statist.* **28** 922–947. MR1792794 https://doi.org/10.1214/aos/1015952005

EL KAROUI, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields* **170** 95–175. MR3748322 https://doi.org/10.1007/s00440-016-0754-9

EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.

FARAWAY, J. J. (1992). Smoothing in adaptive estimation. *Ann. Statist.* **20** 414–427. MR1150352 https://doi.org/10.1214/aos/1176348530

FENG, O. Y., KAO, Y.-C., XU, M. and SAMWORTH, R. J. (2026). Supplement to "Optimal convex M-estimation via score matching." https://doi.org/10.1214/25-AOS2572SUPP

GHOSH, S., IGNATIADIS, N., KOEHLER, F. and LEE, A. (2025). Stein's unbiased risk estimate and Hyvärinen's score matching. arXiv Preprint. Available at arXiv:2502.20123.

GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints: Estimators, Algorithms and Asymptotics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

GUPTA, S., LEE, J. C. H. and PRICE, E. (2023). Finite-sample symmetric mean estimation with Fisher information rate. In *The Thirty Sixth Annual Conference on Learning Theory* 4777–4830. PMLR.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, New York. MR0829458

HANSEN, B. E. (2022). A modern Gauss–Markov theorem. *Econometrica* **90** 1283–1294. MR4436052 https://doi.org/10.3982/ecta19255

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135. MR1766124 https://doi.org/10.1006/jmva.1999.1873

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. And Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I: Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2488795 https://doi.org/10.1002/9780470434697

HYVÄRINEN, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6** 695–709. MR2249836

HYVÄRINEN, A. (2007). Some extensions of score matching. *Comput. Statist. Data Anal.* **51** 2499–2512. MR2338984 https://doi.org/10.1016/j.csda.2006.09.003

JANKOVÁ, J., SHAH, R. D., BÜHLMANN, P. and SAMWORTH, R. J. (2020). Goodness-of-fit testing in high dimensional generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 773–795. MR4112784 https://doi.org/10.1111/rssb.12371

JIN, K. (1990). *Empirical Smoothing Parameter Selection in Adaptive Estimation*. University of California, Berkeley.

JOHNSON, O. (2004). *Information Theory and the Central Limit Theorem*. Imperial College Press, London. MR2109042 https://doi.org/10.1142/9781860945373

JOHNSON, O. and BARRON, A. (2004). Fisher information inequalities and the central limit theorem. *Probab. Theory Related Fields* **129** 391–409. MR2128239 https://doi.org/10.1007/s00440-004-0344-0

JOLICOEUR-MARTINEAU, A., PICHÉ-TAILLEFER, R., DES COMBES, R. T. and MITLIAGKAS, I. (2020). Adversarial score matching and improved sampling for image generation. arXiv Preprint. Available at arXiv:2009.05475.

JONES, M. C. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Ann. Inst. Statist. Math.* **44** 721–727.

KAO, Y.-C., XU, M., FENG, O. Y. and SAMWORTH, R. J. (2024). asm: Optimal convex M-estimation for linear regression via antitonic score matching. R package version 0.2.4. Available at https://CRAN.R-project.org/package=asm.

KAO, Y.-C., XU, M. and ZHANG, C.-H. (2024). Choosing the p in L_p loss: Adaptive rates for symmetric mean estimation. In *The Thirty Seventh Annual Conference on Learning Theory* 2795–2839. PMLR.

KOEHLER, F., HECKETT, A. and RISTESKI, A. (2022). Statistical efficiency of score matching: The view from isoperimetry. arXiv Preprint. Available at arXiv:2210.00726.

LAHA, N. (2021). Adaptive estimation in symmetric location model under log-concavity constraint. *Electron. J. Stat.* **15** 2939–3014. MR4280163 https://doi.org/10.1214/21-ejs1852

LEDERER, J. and OESTING, M. (2023). Extremes in high dimensions: Methods and scalable algorithms. arXiv Preprint. Available at arXiv:2303.04258.

LEI, L. and WOOLDRIDGE, J. (2022). What estimators are unbiased for linear models? arXiv Preprint. Available at arXiv:2212.14185.

LERASLE, M. (2019). Selected topics on robust statistical learning theory. arXiv Preprint. Available at arXiv:1908.10761.

LEY, C. and SWAN, Y. (2013). Stein's density approach and information inequalities. *Electron. Commun. Probab.* **18** no. 7. MR3019670 https://doi.org/10.1214/ECP.v18-2578

LI, G., WEI, Y., CHEN, Y. and CHI, Y. (2024). Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*.

LOH, P.-L. (2021). Scale calibration for high-dimensional robust regression. *Electron. J. Stat.* **15** 5933–5994. MR4355701 https://doi.org/10.1214/21-ejs1936

LYU, S. (2012). Interpretation and generalization of score matching. arXiv Preprint. Available at arXiv:1205.2629.

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400. MR0981457 https://doi.org/10.1214/aos/1176347023

MAMMEN, E. and PARK, B. U. (1997). Optimal smoothing in adaptive location estimation. *J. Statist. Plann. Inference* **58** 333–348. MR1450020 https://doi.org/10.1016/S0378-3758(96)00085-7

MARDIA, K. V., KENT, J. T. and LAHA, A. K. (2016). Score matching estimators for directional distributions. arXiv Preprint. Available at arXiv:1604.08470.

MARONNA, R. A. and YOHAI, V. J. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Z. Wahrsch. Verw. Gebiete* **58** 7–20. MR0635268 https://doi.org/10.1007/BF00536192

NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical Optimization*, 2nd ed. *Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2244940

PARISI, G. (1981). Correlation functions and computer simulations. *Nuclear Phys. B* **180** 378–384. MR0615176 https://doi.org/10.1016/0550-3213(81)90056-0

PARZEN, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74** 105–131. MR0529528

PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499 https://doi.org/10.1214/aos/1176349744

PÖTSCHER, B. M. and PREINERSTORFER, D. (2024). A comment on: "A modern Gauss–Markov theorem". *Econometrica* **92** 913–924. MR4766938 https://doi.org/10.3982/ecta20819

ROBERTS, G. O. and TWEEDIE, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273 https://doi.org/10.2307/3318418

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR0961262

ROCKAFELLAR, R. T. (1997). *Convex Analysis*. *Princeton Landmarks in Mathematics*. Princeton Univ. Press, Princeton, NJ. MR1451876

SAMWORTH, R. and JOHNSON, O. (2004). Convergence of the empirical process in Mallows distance, with an application to bootstrap performance. arXiv Preprint. Available at arXiv:math/0406603.

SAMWORTH, R. J. and SHAH, R. D. (2025). *Modern Statistical Methods and Theory*. Cambridge Univ. Press, Cambridge.

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811 https://doi.org/10.1214/aos/1176350055

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810. MR0663433

SILVERMAN, B. W. (1986). *Density Estimation*. CRC Press, Boca Raton.

SONG, Y. and ERMON, S. (2019). Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inf. Process. Syst.* **32** 11895–11907.

SONG, Y., GARG, S., SHI, J. and ERMON, S. (2020). Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence* 574–584.

SONG, Y. and KINGMA, D. P. (2021). How to train your energy-based models. arXiv Preprint. Available at arXiv:2101.03288.

SONG, Y., SOHL-DICKSTEIN, J., KINGMA, D. P., KUMAR, A., ERMON, S. and POOLE, B. (2021). Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations*.

SRIPERUMBUDUR, B., FUKUMIZU, K., GRETTON, A., HYVÄRINEN, A. and KUMAR, R. (2017). Density estimation in infinite dimensional exponential families. *J. Mach. Learn. Res.* **18** Paper No. 57. MR3687600

STEIN, C. (1956a). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley, CA. MR0084922

STEIN, C. (1956b). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 187–195. Univ. California Press, Berkeley, CA. MR0084921

STONE, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284. MR0362669

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VAN DER VAART, A. W. and WELLNER, J. A. (2021). Stein 1956: Efficient nonparametric testing and estimation. *Ann. Statist.* **49** 1836–1849. MR4319232 https://doi.org/10.1214/21-aos2056

VAN EEDEN, C. (1970). Efficiency-robust estimation of location. *Ann. Math. Statist.* **41** 172–181. MR0263194 https://doi.org/10.1214/aoms/1177697197

VINCENT, P. (2011). A connection between score matching and denoising autoencoders. *Neural Comput.* **23** 1661–1674. MR2839543 https://doi.org/10.1162/NECO_a_00142

YANG, X. and WANG, T. (2024). Multiple-output composite quantile regression through an optimal transport lens. In *The Thirty Seventh Annual Conference on Learning Theory* **247** 5076–5122. PMLR.

YANG, Y., MARTIN, R. and BONDELL, H. (2019). Variational approximations using Fisher divergence. arXiv Preprint. Available at arXiv:1905.05284.

YOHAI, V. J. and MARONNA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7** 258–268. MR0520237

YOUNG, E. H. and SHAH, R. D. (2024). Sandwich boosting for accurate estimation in partially linear models for grouped data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 1286–1311. MR4825003 https://doi.org/10.1093/jrsssb/qkae032

YU, M., GUPTA, V. and KOLAR, M. (2020). Simultaneous inference for pairwise graphical models with generalized score matching. *J. Mach. Learn. Res.* **21** Paper No. 91. MR4119159

YU, S., DRTON, M. and SHOJAIE, A. (2022). Generalized score matching for general domains. *Inf. Inference* **11** 739–780. MR4474347 https://doi.org/10.1093/imaiai/iaaa041

ZOU, H. and YUAN, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** 1108–1126. MR2418651 https://doi.org/10.1214/07-AOS507