Journal of Machine Learning Research 24 (2023) 1-112 Submitted 9/22; Revised 5/23; Published 7/23

Tractable and Near-Optimal Adversarial Algorithms for Robust Estimation in Contaminated Gaussian Models

Ziyue Wang ZW245@STAT.RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, New Jersey 08854, USA

Zhiqiang Tan ZTAN@STAT.RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, New Jersey 08854, USA

Editor: Xiaotong Shen

Abstract

Consider the problem of simultaneous estimation of location and variance matrix under Huber's contaminated Gaussian model. First, we study minimum f-divergence estimation at the population level, corresponding to a generative adversarial method with a non-parametric discriminator and establish conditions on f-divergences which lead to robust estimation, similarly to robustness of minimum distance estimation. More importantly, we develop tractable adversarial algorithms with simple spline discriminators, which can be defined by nested optimization such that the discriminator parameters are determined by maximizing a concave objective function given the current generator. The proposed methods are shown to achieve minimax optimal rates or near-optimal rates depending on the f-divergence and the penalty used. This is the first time such near-optimal error rates are established for adversarial algorithms with linear discriminators under Huber's contamination model. We present simulation studies to demonstrate advantages of the proposed methods over classic robust estimators, pairwise methods, and a generative adversarial method with neural network discriminators.

Keywords: f-divergence, generative adversarial algorithm, Huber's contamination model, minimum divergence estimation, penalized estimation, robust location and scatter estimation

1. Introduction

Consider Huber's contaminated Gaussian model (Huber, 1964): independent observations $X_1, \ldots, X_n$ are obtained from $P_\epsilon = (1-\epsilon)\mathrm{N}(\mu^*, \Sigma^*) + \epsilon Q$, where $\mathrm{N}(\mu^*, \Sigma^*)$ is a p-dimensional Gaussian distribution with mean vector μ^* and variance matrix Σ^*, Q is a probability distribution for contaminated data, and ϵ is a contamination fraction. Our goal is to estimate the Gaussian parameters (μ^*, Σ^*), without any restriction on Q for a small ϵ. This allows both outliers located in areas with vanishing probabilities under $\mathrm{N}(\mu^*, \Sigma^*)$ and other contaminated observations in areas with non-vanishing probabilities under $\mathrm{N}(\mu^*, \Sigma^*)$. We focus on the setting where the dimension p is small relatively to the sample size n, and no sparsity assumption is placed on Σ^* or its inverse matrix. The latter, Σ^{*-1}, is called the precision

matrix and is of particular interest in Gaussian graphical modeling. In the low-dimensional setting, estimation of Σ^* and Σ^{*-1} can be treated as being equivalent.

There is a vast literature on robust statistics (e.g., Huber and Ronchetti, 2009; Maronna et al., 2019) In particular, the problem of robust estimation from contaminated Gaussian data has been extensively studied, and various interesting methods and results have been obtained recently. Under Huber's contamination model above, while the bulk of the data are still Gaussian distributed, a challenge is that the contamination status of each observation is hidden, and the contaminated data may be arbitrarily distributed. In this sense, this problem should be distinguished from various related problems, including multivariate scatter estimation for elliptical distributions as in Tyler (1987) and estimation in Gaussian copula graphical models as in Liu et al. (2012) and Xue and Zou (2012), among others. For motivation and comparison, we discuss below several existing approaches directly related to our work.

Existing work. As suggested by the definition of variance matrix Σ^*, a numerically simple method, proposed in Öllerer and Croux (2015) and Tarr et al. (2016), is to apply a robust covariance estimator for each pair of variables, for example, based on robust scale and correlation estimators, and then assemble those estimators into an estimated variance matrix $\hat{\Sigma}$. These pairwise methods are naturally suitable for both Huber's contamination model and the cellwise contamination model where the components of a data vector can be contaminated independently, each with a small probability ϵ. For various choices of the correlation estimator, such as the transformed Kendall's τ and Spearman's ρ estimator, this method is shown in Loh and Tan (2018) to achieve, in the maximum norm $\|\hat{\Sigma}-\Sigma^*\|_{\max}$, the minimax error rate $\epsilon+\sqrt{\log(p)/n}$ under cellwise contamination and Huber's contamination model. However, because a transformed correlation estimator is used, the variance matrix estimator in Loh and Tan (2018) may not be positive semidefinite (Öllerer and Croux, 2015). Moreover, this approach seems to rely on the availability of individual elements of Σ^* as pairwise covariances and generalization to other multivariate models can be difficult. In our numerical experiments, such pairwise methods have relatively poor performance when contaminated data are not easily separable from the uncontaminated marginally, especially with nonnegligible ϵ.

For location and scatter estimation under Huber's contamination model, Chen et al. (2018) showed that the minimax error rates in the L_2 and operator norm, $\|\hat{\mu}-\mu^*\|_2$ and $\|\hat{\Sigma}-\Sigma^*\|_{\mathrm{op}}$, are $\epsilon+\sqrt{p/n}$ and attained by maximizing Tukey's half-space depth (Tukey, 1975) and a matrix depth function, which is also studied in Zhang (2002) and Paindaveine and Van Bever (2018). Both depth functions, defined through minimization of certain discontinuous objective functions, are in general difficult to compute, and maximization of these depth functions is also numerically intractable. Subsequently, Gao et al. (2019) and Gao et al. (2020) exploited a connection between depth-based estimators and generative adversarial nets (GANs) (Goodfellow et al., 2014), and proposed robust location and scatter estimators in the form of GANs. These estimators are also proved to achieve the minimax error rates in the L_2 and operator norms under Huber's contamination model. More recent work in this direction includes Zhu et al. (2020), Wu et al. (2020), and Liu and Loh (2022).

GANs are a popular approach for learning generative models, with numerous impressive applications (Goodfellow et al., 2014). In the GAN approach, a generator is defined to transform white noises into fake data, and a discriminator is then employed to distinguish

between the fake and real data. The generator and discriminator are trained through minimax optimization with a certain objective function. For GANs used in Gao et al. (2019) and Gao et al. (2020), the generator is defined by the Gaussian model and the discriminator is a multi-layer neural network with sigmoid activations in the top and bottom layers. Hence the discriminator can be seen as logistic regression with the "predictors" defined by the remaining layers of the neural network. The GAN objective function, usually taken to the log-likelihood function in the classification of fake and real data, is more tractable than discontinuous depth functions, but remains nonconvex in the discriminator parameters and nonconcave in the generator parameters. Training such GANs is challenging through nonconvex-nonconcave minimax optimization (see, for example, Farnia and Ozdaglar, 2020 and Jin et al., 2020.)

There is also an interesting connection between GANs and minimum divergence (or distance) (MD) estimation, which has been traditionally studied for robust estimation (Donoho and Liu, 1988; Lindsay, 1994; Basu and Lindsay, 1994). A prominent example is minimum Hellinger distance estimation (Beran, 1977; Tamura and Boos, 1986). In fact, as shown in f-GANs (Nowozin et al., 2016), various choices of the objective function in GANs can be derived from variational lower bounds of f-divergences between the generator and real data distributions. Familiar examples of f-divergences include the Kullback–Leibler (KL), squared Hellinger divergences, and the total variation (TV) distance (Ali and Silvey, 1966; Csiszár, 1967). In particular, using the log-likelihood function in optimizing the discriminator leads to a lower bound of the Jensen–Shannon (JS) divergence for the generator. Furthermore, the lower bound becomes tight if the discriminator class is sufficiently rich (to include the nonparametrically optimal discriminator given any generator). In this sense, f-GANs can be said to nearly implement minimum f-divergence estimation, where the parameters are estimated by minimizing an f-divergence between the model and data distributions. However, this relationship is *only approximate and suggestive*, because even a class of neural network discriminators may not be nonparametrically rich with population data. A similar issue can also be found in the previous studies, where minimum Hellinger estimation and related methods require a smoothed density function of sample data. This approach is impractical for multivariate continuous data.

In addition to MD estimation mentioned above, two other methods of MD estimation have also been studied for robust estimation both in general parametric models and in multivariate Gaussian models. The two methods are defined by minimization of power density divergences (also called β-divergences) (Basu et al., 1998; Miyamura and Kano, 2006) and that of γ-divergences (Windham, 1995; Fujisawa and Eguchi, 2008; Hirose et al., 2017). See Jones et al. (2001) for a comparison of these two methods. In contrast with f-divergences, these two divergences can be evaluated without requiring smooth density estimation from sample data, and hence the corresponding MD estimators can be computed by standard optimization algorithms. To our knowledge, error bounds have not been formally derived for these methods under Huber's contaminated Gaussian model.

Various methods based on iterative pruning or convex programming have been studied with provable error bounds for robust estimation in Huber's contaminated Gaussian model (Lai et al., 2016; Balmand and Dalalyan, 2015; Diakonikolas et al., 2019). These methods either handle scatter estimation after location estimation sequentially in two stages, or resort to using normalized differences of pairs with mean zero for scatter estimation.

Our work. We propose and study adversarial algorithms with linear spline discriminators, and establish various error bounds for simultaneous location and scatter estimation under Huber's contaminated Gaussian model. Two distinct types of GANs are exploited. The first one is logit f-GANs (Tan et al., 2019), which corresponds to a specific choice of f-GANs with the objective function formulated as a negative loss function for logistic regression (or equivalently a density ratio model between fake and real data) when training the discriminator. The second is hinge GAN (Lim and Ye, 2017; Zhao et al., 2017), where the objective function is taken to be the negative hinge loss function when training the discriminator. The hinge objective can be derived from a variational lower bound of the total variation distance (Nguyen et al., 2010; Tan et al., 2019), but cannot be deduced as a special case of the f-GAN objective even though the total variation is also an f-divergence. See Remark 4. In addition, we allow two-objective GANs, including the log$D$ trick in Goodfellow et al. (2014), where two objective functions are used, one for updating the discriminator and the other for updating the generator.

As a major departure from previous studies of GANs, our methods use a simple linear class of spline discriminators, where the basis functions consist of univariate truncated linear functions (or ReLUs shifted) at 5 fixed knots and the pairwise products of such univariate functions. For hinge GAN and certain logit f-GANs including those based on the reverse KL (rKL) and JS divergences, the objective function is concave in the discriminator. By the linearity of the spline class, the objective function is then concave in the spline coefficients. Hence our hinge GAN and logit f-GAN methods involve maximization of a concave function when training the spline discriminator for any fixed generator. In contrast with nonconvex-nonconcave minimax optimization for GANs with neural network discriminators (Gao et al., 2019, 2020), the concavity of the inner optimization for the discriminator contributes to both the numerical tractability and theoretical analysis for our GAN methods. See Remarks 1, 2, 14 and 17. While the optimization for the generator remains nonconvex in our methods, such a single nonconvex optimization is usually more tractable than nonconvex-nonconcave minimax optimization.

In spite of the limited capacity of the spline discriminators, we establish various error bounds for our location and scatter estimators, depending on whether the hinge-GAN or logit f-GAN is used and whether an L_1 or L_2 penalty is incorporated when training the discriminator. See Table 1 for a summary of existing and our error rates in scatter estimation. Our L_1 penalized hinge GAN method achieves the minimax error rate $\epsilon + \sqrt{\log(p)/n}$ in the maximum norm. Our L_2 penalized hinge GAN method achieves the error rate $\epsilon\sqrt{p} + \sqrt{p/n}$, whereas the minimax error rate is $\epsilon + \sqrt{p/n}$, in the $p^{-1/2}$-Frobenius norm. While this might indicate the price paid for maintaining the convexity in training the discriminator, our error rate reduces to the same order $\sqrt{p/n}$ as the minimax error rate provided that ϵ is sufficiently small, $\epsilon = O(\sqrt{1/n})$, such that the contamination error term $\epsilon\sqrt{p}$ is dominated by the sampling variation term $\sqrt{p/n}$ up to a constant factor. To our knowledge, such near-optimal error rates were previously inconceivable for adversarial algorithms with linear discriminators in robust estimation. Moreover, the error rates for our logit f-GAN methods exhibit a square-root dependency on the contamination fraction ϵ, instead of a linear dependency for our hinge GAN methods. This shows, for the first time, some theoretical advantage of hinge GAN over logit f-GANs, although comparative performances of these methods may vary in practice, depending on specific settings.

	Error	Error rate	Computation
OC15, TMW15, LT18	$\|\hat{\Sigma} - \Sigma^*\|_{\max}$	$\epsilon + \sqrt{\log(p)/n}$	Non-iterative computation
CGR18	$\|\hat{\Sigma} - \Sigma^*\|_{\text{op}}$	$\epsilon + \sqrt{p/n}$	Minimax optimization with zero-one discriminators
DKKLMS19	$\|\hat{\Sigma} - \Sigma^*\|_{\text{op}}$	ϵ, provided $\epsilon \geq p/\sqrt{n}$ up to log factors	convex optimization
GYZ20	$\|\hat{\Sigma} - \Sigma^*\|_{\text{op}}$	$\epsilon + \sqrt{p/n}$	Minimax optimization with neural network discriminators
L_1 logit f-GAN (Theorem 11)	$\|\hat{\Sigma} - \Sigma^*\|_{\max}$	$\sqrt{\epsilon} + \sqrt{\log(p)/n}$	
L_1 hinge GAN (Theorem 15)	$\|\hat{\Sigma} - \Sigma^*\|_{\max}$	$\epsilon + \sqrt{\log(p)/n}$	Nested or Minimax optimization with an objective function concave
L_2 logit f-GAN (Theorem 12)	$p^{-\frac{1}{2}}\|\hat{\Sigma} - \Sigma^*\|_{\text{F}}$	$\sqrt{\epsilon} + \sqrt{p/n}$, provided $\epsilon \leq 1/p$ up to a constant factor	in linear spline discriminators
L_2 hinge GAN (Theorem 16)	$p^{-\frac{1}{2}}\|\hat{\Sigma} - \Sigma^*\|_{\text{F}}$	$\epsilon\sqrt{p} + \sqrt{p/n}$	

Table 1: Comparison of existing and proposed methods. OC15, TMW15, LT18 refer to methods and theory in Öllerer and Croux (2015), Tarr et al. (2016), Loh and Tan (2018); CGR18, DKKLMS19, and GYZ20 refer to, respectively, Chen et al. (2018), Diakonikolas et al. (2019), and Gao et al. (2020).

To facilitate and complement our sample analysis, we provide error bounds for the population version of hinge GAN or logit f-GANs with nonparametric discriminators, that is, minimization of the exact total variation or f-divergence at the population level. From Theorem 6, population minimum TV or f-divergence estimation under a simple set of conditions on f (Assumption 1) leads to errors of order $O(\epsilon)$ or $O(\sqrt{\epsilon})$ respectively under Huber's contamination model. Assumption 1 allows the reverse KL, JS, reverse χ^2, and squared Hellinger divergences, but excludes the mixed KL divergence, χ^2 divergence, and, as reassurance, the KL divergence which corresponds to maximum likelihood estimation and is known to be non-robust. Hence certain (but not all) minimum f-divergence estimation achieves robustness under Huber's contamination model or an ϵ TV-contaminated neighborhood. Such robustness is identified for the first time for minimum f-divergence estimation, and is related to, but distinct from, robustness of minimum distance estimation under ϵ contaminated neighborhood with respect to the same distance (Donoho and Liu, 1988). See Remark 9 for further discussion. The population error bounds in the L_2 and $p^{-1/2}$-Frobenius norms are independent of p and hence tighter than the corresponding ϵ terms in our sample error bounds for both hinge GAN and logit f-GAN. These gaps can be attributed to the use of nonparametric versus spline discriminators.

Remarkably, our population analysis also sheds light on the comparison of our sample results and those in Gao et al. (2020). On one hand, another set of conditions (Assumption 2), in addition to Assumption 1, are required in our sample analysis of logit f-GANs with spline discriminators. On the other hand, GANs used in Gao et al. (2020) can be recast as logit f-GANs with neural network discriminators (see Section 5.2). But minimax error rates are shown to be achieved in Gao et al. (2020) for an f-divergence (for example, the mixed KL divergence) which, let alone Assumption 2, does not even satisfy Assumption 1 used in our analysis to show robustness of minimum f-divergence estimation. The main

reason for this discrepancy is that the neural network discriminator in Gao et al. (2020) is directly constrained to be of order $\epsilon + \sqrt{p/n}$ in the log odds, which considerably simplifies the proofs of rate-optimal robust estimation. In contrast, our methods use linear spline discriminators (with penalties independent of ϵ), and our proofs of robust estimation need to carefully tackle various technical difficulties due to the simple design of our methods. See Figure 2(b) for an illustration of non-robustness by minimization of the mixed KL divergence, and Section 5.2 for further discussion on this subtle issue in Gao et al. (2020).

Notation. For a vector $a = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$, we denote by $\|a\|_1 = \sum_{i=1}^p |a_i|$, $\|a\|_\infty = \max_{1 \le i \le p} |a_i|$, and $\|a\|_2 = (\sum_{i=1}^p a_i^2)^{1/2}$ the L_1 norm, L_∞ norm, and L_2 norm of a, respectively. For a matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, we define the element-wise maximum norm $\|A\|_{\max} = \max_{1 \le i \le m, 1 \le j \le n} |a_{ij}|$, the Frobenius norm $\|A\|_{\mathrm{F}} = (\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2)^{1/2}$, the vectorized L_1 norm $\|A\|_{1,1} = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|$, the operator norm $\|A\|_{\mathrm{op}} = \sup_{\|x\|_2 \le 1} \|Ax\|_2$, and the L_∞-induced operator norm $\|A\|_\infty = \sup_{\|x\|_\infty \le 1} \|Ax\|_\infty$. For a square matrix A, we write $A \succeq 0$ to indicate that A is positive semidefinite. The tensor product of vectors a and b is denoted by $a \otimes b$, and the vectorization of matrix A is denoted by $\mathrm{vec}(A)$. The cumulative distribution function of the standard normal distribution is denoted by $\Phi(x)$, and the Gaussian error function is denoted by $\mathrm{erf}(x)$.

2. Numerical illustration

We illustrate the performance of our JS logit f-GAN and existing methods, with two samples of size 20000 from a 100-dimensional Huber's contaminated Gaussian distribution with $\epsilon = 5\%$ and 20%, based on a Toeplitz variance matrix and the first Cauchy contamination Q in Section 6.2. Figure 1 shows the 95% Gaussian ellipses for two selected coordinates, using the estimated location vectors and variance matrices except for Tyler's M-estimator (Tyler, 1987), Kendalls's τ with MAD (Loh and Tan, 2018), and Spearman's ρ with Q_n-estimator (Öllerer and Croux, 2015) where the locations are set to the true means. The performances of our rKL logit f-GAN and hinge GAN are close to that of JS logit f-GAN. See Figure 5 for illustration based on the second contamination in Section 6.2.

Among the methods shown in Figure 1, the JS logit f-GAN gives an estimated ellipse that is closest to the truth, followed with small but noticeable differences by the JS-GAN (Gao et al., 2020). The MCD (Rousseeuw, 1985) performs among the best when $\epsilon = 5\%$ but deteriorates considerably when ϵ increases to 20%. The remaining three methods, Kendall's τ with MAD, Spearman's ρ with Q_n-estimator, and Tyler's M-estimator show much less satisfactory performance. The estimated distributions from these methods are dragged towards the corner contamination cluster.

The relatively poor performance of the pairwise methods, Kendall's τ with MAD and Spearman's ρ with Q_n-estimator, may be explained by the fact that as shown by the marginal histograms in Figure 1, the data in each coordinate are one-sided heavy-tailed, but no obvious outliers can be seen marginally. The correlation estimates from Kendall's τ and Spearman's ρ tend to be inaccurate even after sine transformations, especially with nonnegligible $\epsilon = 20\%$. In contrast, our GAN methods and JS-GAN, as well as MCD in the case of $\epsilon = 5\%$, are capable of capturing higher dimensional information so that the impact of contamination is limited to various extents.

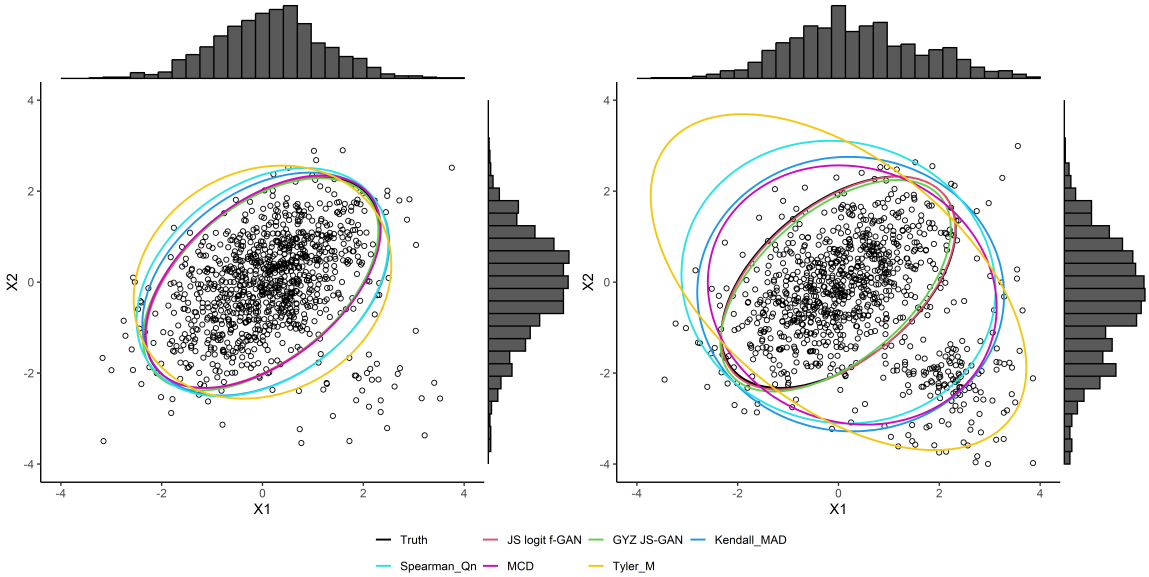

Figure 1: The estimated 95% Gaussian ellipses and marginal histograms for two selected coordinates, from contaminated data based on the first Cauchy contamination in Section 6.2 with $\epsilon = 5\%$ (left) or 20% (right). For visibility, the data points are truncated to $(-4, 4)$ on each axis; see Appendix A for untruncated plots.

3. Background: Adversarial algorithms

We review various adversarial algorithms (or GANs), which are exploited by our methods for robust location and scatter estimation. To focus on main ideas, the algorithms are stated in their population versions, where the underlying data distribution P_* is involved instead of the empirical distribution P_n. Let $\{P_\theta : \theta \in \Theta\}$ be a statistical model and $\{h_\gamma : \gamma \in \Gamma\}$ be a function class, where P_θ is called a generator and h_γ a discriminator. In our study, P_θ is a multivariate Gaussian distribution $\mathrm{N}(\mu, \Sigma)$, and h_γ is a pairwise spline function which is specified later in Section 4.2.

For a convex function $f : (0, \infty) \to \mathbb{R}$, the f-divergence between the distributions P_* and P_θ with density functions p_* and p_θ is

$$D_f(P_* \| P_\theta) = \int f\left(\frac{p_*(x)}{p_\theta(x)}\right) \mathrm{d}P_\theta.$$

For example, taking $f(t) = t \log t$ yields the Kullback–Liebler (KL) divergence $D_{\mathrm{KL}}(P_* \| P_\theta)$. The logit f-GAN (Tan et al., 2019) is defined by solving the minimax program

$$\min_{\theta \in \Theta} \max_{\gamma \in \Gamma} \; K_f(P_*, P_\theta; h_\gamma), \tag{1}$$

where

$$\begin{aligned} K_f(P_*, P_\theta; h) &= \mathrm{E}_{P_*} f'(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_\theta} f^{\#}(e^{h(x)}) \\ &= \mathrm{E}_{P_*} f'(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_\theta} \left\{ \mathrm{e}^{h(x)} f'(\mathrm{e}^{h(x)}) - f(e^{h(x)}) \right\}. \end{aligned}$$

Throughout, $f^{\#}(t) = tf'(t) - f(t)$ and f' denotes the derivative of f. A motivation for this method is that the objective K_f is a nonparametrically tight, lower bound of the f-divergence (Tan et al., 2019, Proposition S1): for each θ, it holds that for any function h,

$$K_f(P_*, P_\theta; h) \le D_f(P_* \| P_\theta), \tag{2}$$

where the equality is attained at $h_{*\theta}(x) = \log\{p_*(x)/p_\theta(x)\}$, the log density ratio between P_* and P_θ or equivalently the log odds for classifying whether a data point x is from P_* or P_θ. There are two choices of f of particular interest. Taking $f(t) = t\log t - (t+1)\log(t+1) + \log 4$ leads the Jensen–Shannon (JS) divergence, $D_{\mathrm{JS}}(P_*\|P_\theta) = D_{\mathrm{KL}}(P_*\|(P_*+P_\theta)/2) + D_{\mathrm{KL}}(P_\theta\|(P_*+P_\theta)/2)$, and the objective function

$$K_{\mathrm{JS}}(P_*, P_\theta; h) = -\mathrm{E}_{P_*}\log(1+\mathrm{e}^{-h(x)}) - \mathrm{E}_{P_\theta}\log(1+\mathrm{e}^{h(x)}) + \log 4,$$

which is, up to a constant, the expected log-likelihood for logistic regression with log odds function $h(x)$. For $K_f = K_{\mathrm{JS}}$, program (1) corresponds to the original GAN (Goodfellow et al., 2014) with discrimination probability sigmoid$(h(x))$. Taking $f(t) = -\log t$ leads to the reverse KL divergence $D_{\mathrm{rKL}}(P_*\|P_\theta) = D_{\mathrm{KL}}(P_\theta\|P_*)$ and the objective function

$$K_{\mathrm{rKL}}(P_*, P_\theta; h) = 1 - \mathrm{E}_{P_*}\mathrm{e}^{-h(x)} - \mathrm{E}_{P_\theta}h(x),$$

which is the negative calibration loss for logistic regression in Tan (2020).

The objective K_f with fixed θ can be seen as a proper scoring rule reparameterized in terms of the log odds function $h(x)$ for binary classification (Tan and Zhang, 2022). Replacing K_f in (1) by the negative hinge loss (which is not a proper scoring rule) leads to

$$\min_{\theta\in\Theta}\max_{\gamma\in\Gamma}\ K_{\mathrm{HG}}(P_*, P_\theta; h_\gamma), \tag{3}$$

where

$$K_{\mathrm{HG}}(P_*, P_\theta; h) = \mathrm{E}_{P_*}\min(1, h(x)) + \mathrm{E}_{P_\theta}\min(1, -h(x)).$$

This method is related to the geometric GAN described later in (7) and will be called hinge GAN. By Nguyen et al. (2009) or Proposition 5 in Tan et al. (2019), the objective K_{HG} is a nonparametrically tight, lower bound of the total variation distance scaled by 2: for each θ, it holds that for any function $h(x)$,

$$K_{\mathrm{HG}}(P_*, P_\theta; h) \le 2D_{\mathrm{TV}}(P_*\|P_\theta), \tag{4}$$

where the equality is attained at $h_{*\theta}(x) = \mathrm{sign}(p_*(x) - p_\theta(x))$, and $D_{\mathrm{TV}}(P_*\|P_\theta) = \int |p_*(x) - p_\theta(x)|/2\,\mathrm{d}x$. The objectives K_f and K_{HG}, with fixed θ, represent two types of loss functions for binary classification. See Buja et al. (2005) and Nguyen et al. (2009) for further discussions about loss functions and scoring rules.

The preceding programs, (1) and (3), are defined as minimax optimization, each with a single objective function. There are also adversarial algorithms, which are formulated as

alternating optimization with two objective functions (see Remark 1). For example, GAN with the $\log D$ trick in Goodfellow et al. (2014) is defined by solving

$$\begin{cases} \max\limits_{\gamma\in\Gamma} \ K_{\mathrm{JS}}(P_*, P_\theta; \gamma) & \text{with } \theta \text{ fixed,} \\ \min\limits_{\theta\in\Theta} \ \mathrm{E}_{P_\theta} \log(1 + \mathrm{e}^{-h_\gamma(x)}) & \text{with } \gamma \text{ fixed.} \end{cases} \tag{5}$$

The second objective is introduced mainly to overcome vanishing gradients in θ when the discriminator is confident. The calibrated rKL-GAN (Huszár, 2016; Tan et al., 2019) is defined by solving

$$\begin{cases} \max\limits_{\gamma\in\Gamma} \ K_{\mathrm{JS}}(P_*, P_\theta; \gamma) & \text{with } \theta \text{ fixed,} \\ \min\limits_{\theta\in\Theta} \ -\mathrm{E}_{P_\theta} h_\gamma(x) & \text{with } \gamma \text{ fixed.} \end{cases} \tag{6}$$

The two objectives are chosen to stabilize gradients in both θ and γ during training. The geometric GAN in Lim and Ye (2017) or, equivalently, the energy-based GAN in Zhao et al. (2017) as shown in Tan et al. (2019), is defined by solving

$$\begin{cases} \max\limits_{\gamma\in\Gamma} \ K_{\mathrm{HG}}(P_*, P_\theta; \gamma) & \text{with } \theta \text{ fixed,} \\ \min\limits_{\theta\in\Theta} \ -\mathrm{E}_{P_\theta} h_\gamma(x) & \text{with } \gamma \text{ fixed.} \end{cases} \tag{7}$$

Interestingly, the second line in (6) or (7) involves the same objective $-\mathrm{E}_{P_\theta} h_\gamma(x)$, which can be equivalently replaced by $K_{\mathrm{rKL}}(P_*, P_\theta; h_\gamma)$ because γ and hence h_γ are fixed.

Remark 1 *We discuss precise definitions for a solution to a minimax problem such as (1) or (3), and a solution to an alternating optimization problem such as (5)–(7). For an objective function $K(\theta, \gamma)$, we say that $(\hat\theta, \hat\gamma)$ is a solution to*

$$\min_\theta \max_\gamma \ K(\theta, \gamma), \tag{8}$$

if $K(\hat\theta, \hat\gamma) = \max_\gamma K(\hat\theta, \gamma) \le \max_\gamma K(\theta, \gamma)$ for any θ. In other words, we treat (8) as nested optimization: $\hat\theta$ is a minimizer of $K(\theta, \hat\gamma_\theta)$ as a function of θ and $\hat\gamma = \hat\gamma_{\hat\theta}$, where $\hat\gamma_\theta$ is a maximizer of $K(\theta, \gamma)$ for fixed θ. This choice is directly exploited in both numerical implementation and theoretical analysis of our methods later. For two objective functions $K_1(\theta, \gamma)$ and $K_2(\theta, \gamma)$, we say that $(\hat\theta, \hat\gamma)$ is a solution to the alternating optimization problem

$$\begin{cases} \max\limits_{\gamma} \ K_1(\theta, \gamma) & \textit{with } \theta \textit{ fixed,} \\ \min\limits_{\theta} \ K_2(\theta, \gamma) & \textit{with } \gamma \textit{ fixed,} \end{cases} \tag{9}$$

if $K_1(\hat\theta, \hat\gamma) = \max_\gamma K_1(\hat\theta, \gamma)$ and $K_2(\hat\theta, \hat\gamma) = \min_\theta K_2(\theta, \hat\gamma)$. In the special case where $K_1(\theta, \gamma) = K_2(\theta, \gamma)$, denoted as $K(\theta, \gamma)$, a solution $(\hat\theta, \hat\gamma)$ to (9) is also called a Nash equilibrium of $K(\theta, \gamma)$, satisfying $K(\hat\theta, \hat\gamma) = \max_\gamma K(\hat\theta, \gamma) = \min_\theta K(\theta, \hat\gamma)$. It can be shown that a Nash equilibrium of $K(\theta, \gamma)$ is equivalently a solution to both minimax problem (8) and the maximin problem $\max_\gamma \min_\theta \ K(\theta, \gamma)$, similarly treated as nested optimization. For general $K(\theta, \gamma)$, the minimax and maximin solutions may differ from each other, although

Algorithm 1: Gradient descent ascent

Require A GAN objective function $K(\theta, \gamma)$ as in (1) or (3) with P_* replaced by the empirical distribution P_n on real data, initial values (θ_0, γ_0), learning rates (α_d, α_g) for the generator and discriminator, and number of epochs T.

for $t = 1 \dots T$ **do**

Sampling: Generate a sample of fake data to approximate $P_{\theta_{t-1}}$.

Updating: Compute $\gamma_t = \gamma_{t-1} + \alpha_d \nabla_\gamma K(\theta, \gamma)|_{\gamma_{t-1}}$, and $\theta_t = \theta_{t-1} - \alpha_g \nabla_\theta K(\theta, \gamma_t)|_{\theta_{t-1}}$.

end

Algorithm 2: Gradient descent with concave inner optimization

Require A GAN objective function $K(\theta, \gamma)$, which is concave in γ for each fixed θ, initial value θ_0, learning rate α_g for the generator, and number of epochs T.

for $t = 1 \dots T$ **do**

Sampling: Generate a sample of fake data to approximate $P_{\theta_{t-1}}$.

Updating: Compute $\gamma_t = \text{argmax}_\gamma K(\theta_{t-1}, \gamma)$ by a concave optimizer, and $\theta_t = \theta_{t-1} - \alpha_g \nabla_\theta K(\theta, \gamma_t)|_{\theta_{t-1}}$.

end

they coincide and yield a Nash equilibrium by Sion's minimax theorem in the special setting where $K(\theta, \gamma)$ is convex in θ for each γ and concave in γ for each θ. Our definition of single-objective GANs as nested optimization agrees with Jin et al. (2020), where a solution to (8) is called a (global) minimax point of $K(\theta, \gamma)$, and hence should be distinguished from the interpretation of GANs as finding Nash equilibria or modifications (Farnia and Ozdaglar, 2020). On the other hand, our definition of two-objective GANs takes the form of alternating optimization, which is currently needed for our theoretical analysis (Section 4.4). It remains open whether theoretical guarantees can also be developed for two-objective GANs based on nested optimization (even with neural network discriminators).

Remark 2 *Numerically, GANs are often trained using gradient-based algorithms, notably the gradient descent ascent algorithm (GDA) (Algorithm 1), which can only be expected to find local solutions. However, there are subtle issues even in the consideration of local solutions. Formally as in Jin et al. (2020), a point $(\hat{\theta}, \hat{\gamma})$ is said to be a local minimax point of $K(\theta, \gamma)$ if there exist $\delta_0 > 0$ and a function $h(\cdot)$ satisfying $h(\delta) \to 0$ as $\delta \to 0$, such that for any $\delta \in (0, \delta_0]$, the pair $(\hat{\theta}, \hat{\gamma})$ satisfies*

$$K(\hat{\theta}, \hat{\gamma}) = \max_{\gamma : \|\gamma - \hat{\gamma}\| \le h(\delta)} K(\hat{\theta}, \gamma) \le \max_{\gamma : \|\gamma - \hat{\gamma}\| \le h(\delta)} K(\theta, \gamma),$$

for any θ satisfying $\|\theta - \hat{\theta}\| \le \delta$. As shown in Jin et al. (2020), a local minimax point can be characterized via necessary and sufficient conditions based on the gradients and Hessians. But for nonconvex-nonconcave minimax optimization, a global minimax point may be neither a local minimax point nor a stationary point (i.e., with the gradients of K being 0 with respect to θ and γ). This differs markedly from nonconvex optimization

where a global minimum is always a local minimum and, if in the interior of the domain, a stationary point. Nevertheless, our GAN methods in Section 4 are designed such that $K(\theta, \gamma)$ is concave in γ for each fixed θ. In this setting, as noted in Jin et al. (2020), a global minimax point is always a local minimax point, and hence finding a local minimax point through GDA is provably a feasible strategy for finding a global minimax point. Moreover, for our methods with concave $K(\theta, \cdot)$, nested optimization can be directly implemented, as shown in Algorithm 2, using a concave optimizer over γ and gradient descent over θ to find a local minimizer of $L(\theta) = K(\theta, \hat{\gamma}_\theta)$, with $\hat{\gamma}_\theta = \text{argmax}_\gamma K(\theta, \gamma)$. This is a feasible example of gradient descent with max-oracle, for which a performance guarantee is derived in Jin et al. (2020). Based on these observations, our GAN methods with concave $K(\theta, \cdot)$ are numerically more tractable than nonconvex-nonconcave GANs.

Remark 3 *The population f-GAN (Nowozin et al., 2016) is defined by solving*

$$\min_{\theta\in\Theta} \max_{\gamma\in\Gamma} \left\{E_{P_*} T_\gamma(x) - E_{P_\theta} f^*(T_\gamma(x))\right\}, \tag{10}$$

where f^* is the Fenchel conjugate of f, i.e., $f^*(s) = \sup_{t\in(0,\infty)}(st - f(t))$ and T_γ is a function taking values in the domain of f^*. Typically, T_γ is represented as $T_\gamma(x) = \tau_f(h_\gamma(x))$, where $\tau_f : \mathbb{R} \to \text{dom}(f^*)$ is an activation function and $h_\gamma(x)$ take values unrestricted in $\mathbb{R}$. The logit f-GAN corresponds to f-GAN with the specific choice $\tau_f(u) = f'(\mathrm{e}^u)$ by the relationship $f^*(f'(t)) = f^\#(t)$ (Tan et al., 2019). Nevertheless, a benefit of logit f-GAN is that the objective K_f in (1) takes the explicit form of a negative discrimination loss such that $h_\gamma(x)$ can be seen to approximate the log density ratio between P_* and P_θ.

Remark 4 *There is an important difference between hinge GAN and logit f-GAN, although the total variation is also an f-divergence with $f(t) = |t-1|/2$. In fact, taking this choice of f in logit f-GAN (1) yields*

$$\min_{\theta\in\Theta} \max_{\gamma\in\Gamma} \left\{E_{P_*} \text{sign}(h_\gamma(x)) - E_{P_\theta} \text{sign}(h_\gamma(x))\right\}. \tag{11}$$

This is called TV learning and is related to depth-based estimation in Gao et al. (2019). Compared with hinge GAN in (3), program (11) is computationally more difficult to solve. Such a difference also exists in the application of general f-GAN to the total variation. For the total variation distance scaled by 2 with $f(t) = |t-1|$, the conjugate is $f^*(s) = \max(-1, s)$ if $s \le 1$ or ∞ if $s > 1$. If T_γ is specified as $T_\gamma = \min(1, h_\gamma(x))$, then the objective in f-GAN (10) can be shown to be

$$E_{P_*} \min(1, h_\gamma(x)) + \mathrm{E}_{P_\theta} \min(1, \max(-1, -h_\gamma(x))),$$

which in general differs from the negative hinge loss in (3) unless h_γ is upper bounded by 1. If h_γ is specified as $2\,\text{sigmoid}(\tilde{h}_\gamma) - 1$ for a function $\tilde{h}_\gamma$ taking values unrestricted in $\mathbb{R}$, the resulting f-GAN is equivalent to TV-GAN in Gao et al. (2019) defined by solving

$$\min_{\theta\in\Theta} \max_{\gamma\in\Gamma} \left\{E_{P_*} \text{sigmoid}(\tilde{h}_\gamma(x)) - E_{P_\theta} \text{sigmoid}(\tilde{h}_\gamma(x))\right\}. \tag{12}$$

However, solving program (12) is numerically intractable as discussed in Gao et al. (2019).

4. Theory and methods

We propose and study various adversarial algorithms with simple spline discriminators for robust estimation in a multivariate Gaussian model. Assume that $X_1, \ldots, X_n$ are independent observations obtained from Huber's ϵ-contamination model, that is, the data distribution P_* is of the form

$$P_\epsilon = (1-\epsilon)P_{\theta^*} + \epsilon Q, \tag{13}$$

where P_{θ^*} is $\mathrm{N}(\mu^*, \Sigma^*)$ with unknown $\theta^* = (\mu^*, \Sigma^*)$, Q is a probability distribution for contaminated data, and ϵ is a contamination fraction. Both Q and ϵ are unknown and Q can be an arbitrary probability distribution. The dependency of P_ϵ on (θ^*, Q) is suppressed in the notation. Equivalently, the data $(X_1, \ldots, X_n)$ can be represented in a latent model: $(U_1, X_1), \ldots, (U_n, X_n)$ are independent, and U_i is Bernoulli with $\mathrm{P}(U_i = 1) = \epsilon$ and X_i is drawn from P_{θ^*} or Q given $U_i = 0$ or 1 for $i = 1, \ldots, n$.

For theoretical analysis, we consider two choices of the parameter space. The first choice is $\Theta_1 = \{(\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma \succeq 0, \|\Sigma\|_{\max} \le M_1\}$ for a constant $M_1 > 0$. Equivalently, the diagonal elements of Σ is upper bounded by M_1 for $(\mu, \Sigma) \in \Theta_1$. The second choice is $\Theta_2 = \{(\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma \succeq 0, \|\Sigma\|_{\mathrm{op}} \le M_2\}$ for a constant $M_2 > 0$. For simplicity, the dependency of Θ_1 on M_1 or Θ_2 on M_2 is suppressed in the notation. For the second parameter space Θ_2, the minimax rates in the L_2 and operator norms have been shown to be achieved using matrix depth (Chen et al., 2018) and GANs with certain neural network discriminators (Gao et al., 2020).

Our work aims to investigate adversarial algorithms with a simple linear class of spline discriminators for computational tractability, and establish various error bounds for the proposed estimators, including those matching the minimax rates in the maximum norms for the location and scatter estimation over Θ_1, and, provided that $\epsilon\sqrt{n}$ is bounded by a constant (independent of p), the minimax rates in the L_2 and Frobenius norms over Θ_2.

It is worth emphasizing that adversarial algorithms is used in our work to learn the multivariate Gaussian distribution P_{θ^*} with the real data assumed to be from Huber's contaminated Gaussian distribution P_ϵ for some unknown (Q, ϵ), in addition to the unknown parameter θ^*. Hence this differs from the usual theoretical setting where the real data are assumed to be generated purely from the model distribution P_{θ^*}.

4.1 Population analysis with nonparametric discriminators

A distinctive feature of GANs is that they can be motivated as approximations to minimum divergence estimation. For example, if the discriminator class $\{h_\gamma\}$ in (1) is rich enough to include the nonparametrically optimal discriminator such that $\max_{\gamma\in\Gamma} K_f(P_*, P_\theta; h_\gamma) = D_f(P_*\|P_\theta)$ for each θ, then the (population) logit f-GAN amounts to minimizing the f-divergence $D_f(P_*\|P_\theta)$. Similarly, if the discriminator class $\{h_\gamma\}$ in (3) is sufficiently rich, then the (population) hinge GAN amounts to minimizing the total variation $D_{\mathrm{TV}}(P_*\|P_\theta)$.

As a prelude to our sample analysis, Theorem 6 shows that at the population level, minimization of the total variation and certain f-divergences satisfying Assumption 1 achieves robustness under Huber's contamination model, in the sense that the estimation errors are respectively $O(\epsilon)$ and $O(\sqrt{\epsilon})$, uniformly over all possible Q. Hence with sufficiently rich (or nonparametric) discriminators, the population versions of the hinge GAN and certain

Name	Convex $f(t)$	Non-incr. $f(t)$	Concave $f'(t)$	Concave $f'(\mathrm{e}^u)$	Lipschitz $f^{\#}(\mathrm{e}^u)$
Total variation	$(1-t)_+, \lvert t-1\rvert/2$	✓	—	—	—
Reverse KL	$-\log t$	✓	✓	✓	✓
Jensen-Shannon	$t\log t-(t+1)\log(t+1)+\log 4$	✓	✓	✓	✓
Squared Hellinger	$(\sqrt{t}-1)^2$	✓	✓	✓	
Reverse χ^2	$t^{-1}-1$	✓	✓	✓	
KL	$t\log t$		✓	✓	
Mixed KL	$\{(t-1)\log t\}/2$		✓	✓	
χ^2	$(t-1)^2$		✓		

Table 2: Common f-divergences and validity of Assumptions 1 (ii)–(iii) and 2 (i)–(ii). The mixed KL divergence is defined as $D_{\mathrm{mKL}}(P||Q) = D_{\mathrm{KL}}(P||Q)/2 + D_{\mathrm{KL}}(Q||P)/2$.

f-GANs can be said to be robust under Huber's contamination. From Table 2, Assumption 1 is satisfied by the reverse KL, JS, and squared Hellinger divergences, but violated by the KL divergence. Minimization of the KL divergence corresponds to maximum likelihood estimation, which is known to be non-robust under Huber's contamination model.

Assumption 1 *Suppose that $f : (0,\infty) \to \mathbb{R}$ is convex with $f(1) = 0$ and satisfies the following conditions.*

(i) f is twice differentiable with $f''(1) > 0$.

(ii) f is non-increasing.

(iii) f' is concave (i.e., f'' is non-increasing)

See Table 2 for validity of conditions (ii) and (iii) in various f-divergences.

Remark 5 *Given a convex function f with $f(1) = 0$, the same f-divergence D_f can be defined using the convex function $f(t) + c(t-1)$ for any constant $c \in \mathbb{R}$. Hence condition (ii) in Assumption 1 can be relaxed such that f' is upper bounded by a constant. The non-increasingness of f is stated above for ease of interpretation. The other conditions in Assumption 1 and Assumption 2 are not affected by non-unique choices of f.*

Theorem 6 *Let $\Theta_0 = \{(\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma$ is a $p \times p$ variance matrix$\}$.*

(i) Assume that f satisfies Assumption 1. Let $\bar{\theta} = \mathrm{argmin}_{\theta\in\Theta_0} D_f(P_\epsilon||P_\theta)$. If $\sqrt{-2(f''(1))^{-1}f'(1/2)\epsilon} + \epsilon < 1/2$, then for any contamination distribution Q,

$$\|\bar{\mu} - \mu^*\|_2 \le C\|\Sigma^*\|_{\mathrm{op}}^{1/2}\sqrt{\epsilon}, \quad \|\bar{\mu} - \mu^*\|_\infty \le C\|\Sigma^*\|_{\max}^{1/2}\sqrt{\epsilon}, \tag{14}$$

and

$$\|\bar{\Sigma} - \Sigma^*\|_{\mathrm{op}} \le C\|\Sigma^*\|_{\mathrm{op}}\sqrt{\epsilon}, \quad \|\bar{\Sigma} - \Sigma^*\|_{\max} \le C\|\Sigma^*\|_{\max}\sqrt{\epsilon}, \tag{15}$$

where $C > 0$ is a constant depending only on f. The same inequality as in (15) also holds with $\|\bar{\Sigma} - \Sigma^*\|_{\mathrm{op}}$ replaced by $p^{-1/2}\|\bar{\Sigma} - \Sigma^*\|_{\mathrm{F}}$.

(ii) Let $\bar{\theta} = \mathrm{argmin}_{\theta\in\Theta_0} D_{\mathrm{TV}}(P_\epsilon||P_\theta)$. If $\epsilon < 1/4$ then (14) and (15) hold for an absolute constant $C > 0$ with $\sqrt{\epsilon}$ replaced by ϵ throughout.

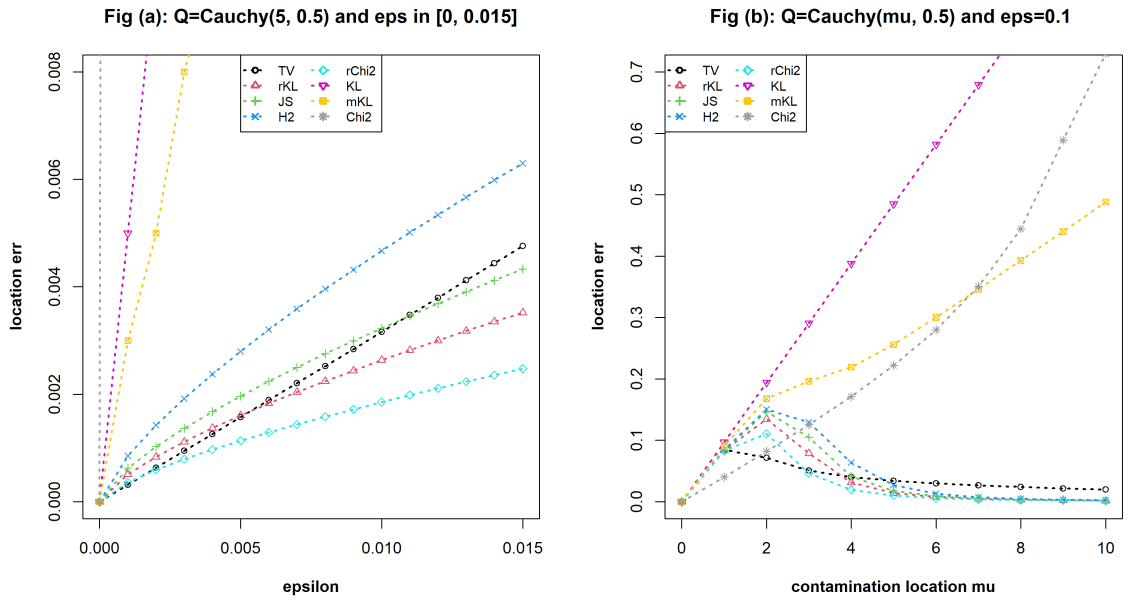

Figure 2: Illustration of robustness of minimum f-divergence location estimation. Figure (a): Location error $|\bar{\mu}-\mu^*|$ against contamination fraction ϵ from 0 to 0.015, with P_{θ^*} being $\mathrm{N}(0,1)$ and contamination Q being $\mathrm{Cauchy}(5,1/2)$ fixed; Figure (b): Location error $|\bar{\mu}-\mu^*|$ against contamination location μ_Q from 0 to 10, with P_{θ^*} being $\mathrm{N}(0,1)$, contamination fraction $\epsilon = 0.1$ fixed, and contamination Q being $\mathrm{Cauchy}(\mu_Q, 1/2)$. The squared Hellinger, reverse χ^2, and mixed KL are denoted by H2, rChi2, and mKL respectively.

Figure 2 provides a simple numerical illustration. From Figure 2(a), the location errors $|\bar{\mu}-\mu^*|$ of minimum divergence estimators corresponding to the four robust f-divergences (reverse KL, JS, squared Hellinger, and reverse χ^2) satisfying Assumption 1 are of shapes in agreement with the order $\sqrt{\epsilon}$ in Theorem 6, whereas those corresponding to TV appear to be linear in ϵ, for ϵ close to 0. For the KL, mixed KL, and χ^2 divergences, which do not satisfy Assumption 1(ii), their corresponding errors quickly increase out of the plotting range, indicating non-robustness of the associated minimum divergence estimation. The differences between robust and non-robust f-divergences are further demonstrated in Figure 2(b). As the contamination location moves farther away, the errors of the robust f-divergences increase initially but then decrease to near 0, whereas those of the non-robust f-divergences appear to increase unboundedly.

Remark 7 *From the proof in Section 7.1, Theorem 6(i) remains valid if $f''(1)$ is replaced by $C_f = \inf_{t\in(0,1]} f''(t)$ in Assumption 1(i) and the definition of $\mathrm{Err}_{f0}(\epsilon)$, and Assumption 1(iii), the concavity of f', is removed. On the other hand, a stronger condition than Assumption 1(iii) is used in our sample analysis: for convex f, the concavity of f' is implied by Assumption 2(i), as discussed in Remark 13.*

Remark 8 *The population bounds in Theorem 6 are more refined than those in our sample analysis later. The population minimizer $\bar{\theta} = (\bar{\mu}, \bar{\Sigma})$ is defined by minimization over the unrestricted space Θ_0 instead of Θ_1 or Θ_2 with the restriction $\|\Sigma\|_{\max} \le M_1$ or $\|\Sigma\|_{\rm op} \le M_2$. The population bounds are also adaptive in that the scaling constants depend directly on the maximum or operator norm of the true variance matrix Σ^*, instead of pre-specified constants M_1 or M_2. Note that the parameter space is also restricted such that $\|\Sigma\|_{\rm op} \le M_2$ and the error bounds depend on M_2 in sample analysis of Gao et al. (2020). Nevertheless, the population bounds share a similar feature as in our sample bounds later: the error bounds in the maximum norms are governed by $\|\Sigma^*\|_{\max}$, which can be much smaller than $\|\Sigma^*\|_{\rm op}$ involved in the error bounds in the operator norm.*

Remark 9 *It is interesting to connect and compare our results with Donoho and Liu (1988), where minimum distance (MD) estimation is studied, that is, minimization of a proper distance $D(P, P_\theta)$ satisfying the triangle inequality. For minimum TV estimation, let $\bar{\theta}_P = (\bar{\mu}_P, \bar{\Sigma}_P) = \mathrm{argmin}_\theta D_{\rm TV}(P\|P_\theta)$. For location estimation, define*

$$b(\epsilon) = \sup_{P: D_{\rm TV}(P\|P_{\theta^*}) \le \epsilon} \|\bar{\mu}_P - \mu^*\|_2, \qquad b_0(\epsilon) = \sup_{\theta: D_{\rm TV}(P_\theta\|P_{\theta^*}) \le \epsilon} \|\mu - \mu^*\|_2,$$

which are called the bias distortion curve and the gauge function. Scatter estimation can be discussed in a similar manner. For a general family $\{P_\theta\}$, the first half in our proof of Theorem 6(ii) shows that for any P satisfying $D_{\rm TV}(P\|P_{\theta^*}) \le \epsilon$, we have $D_{\rm TV}(P_{\bar{\theta}_P}\|P_{\theta^*}) \le 2\epsilon$. This implies a bound similar to Proposition 5.1 in Donoho and Liu (1988):

$$b(\epsilon) \le b_0(2\epsilon). \tag{16}$$

For the multivariate Gaussian family $\{P_\theta\}$, the second half in our proof of Theorem 6(ii) derives an explicit upper bound on $b_0(\epsilon)$ provided that $2\epsilon \le a$ for a constant $a \in [0, 1/2)$:

$$b_0(2\epsilon) \le S_{1,a} \|\Sigma^*\|_{\rm op}^{1/2}(2\epsilon),$$

where $S_{1,a} = \{\Phi'(\Phi^{-1}(1/2 + a))\}^{-1}$. Combining the preceding inequalities yields $b(\epsilon) \le C\|\Sigma^*\|_{\rm op}^{1/2}\epsilon$ in Theorem 6(ii), with $C = 2S_{1,a}$. In addition, Proposition 5.1 in Donoho and Liu (1988) gives the same bound as (16) for MD estimation using certain other distances $D(P, P_\theta)$, including the Hellinger distance, where the MD functional $\bar{\theta}_P = (\bar{\mu}_P, \bar{\Sigma}_P)$ is defined as $\mathrm{argmin}_\theta D(P, P_\theta)$, and $b(\epsilon)$ and $b_0(\epsilon)$ are defined with $D_{\rm TV}(P\|P_{\theta^*})$ replaced by $D(P, P_\theta)$. The distances used in defining the MD functional and the contamination neighborhood are tied to each other. Hence, except for minimum TV estimation, our setting differs from Donoho and Liu (1988) in studying different choices of minimum f-divergence estimation over the same Huber's contamination neighborhood.

Remark 10 *We briefly comment on how our result is related to breakdown points in robust statistics (Huber and Ronchetti, 2009, Section 1.4). For estimating μ^*, the population breakdown point of a functional $T = T(P)$ can be defined as $\sup\{\epsilon : b_T(\epsilon) < \infty\}$, where $b_T(\epsilon) = \sup_{P: D_{\rm TV}(P\|P_{\theta^*}) \le \epsilon} \|T(P) - \mu^*\|_2$. Scatter estimation can be discussed in a similar manner. For T defined from minimum TV estimation, Theorem 6(ii) shows that if $\epsilon < 1/4$, then $b_T(\epsilon) \le C\|\Sigma^*\|_{\rm op}^{1/2}\epsilon$, as noted in Remark 9. This not only provides an explicit bound*

on $b_T(\epsilon)$, but also implies that the population breakdown point is at least $1/4$ for minimum TV estimation. Similar implications can be obtained from Theorem 6(i) for minimum f-divergence estimation. For T defined from minimum rKL divergence estimation, Theorem 6(i) shows that if $2\sqrt{\epsilon} + \epsilon < 1/2$, then $b_T(\epsilon) \le C\|\Sigma^*\|_{\mathrm{op}}^{1/2}\sqrt{\epsilon}$, and hence the population breakdown point is at least 0.051. While these estimates of breakdown points can potentially be improved, our population analysis as well as sample analysis in the subsequent sections focus on deriving quantitative error bounds in terms of sufficiently small ϵ and some scaling constants free of ϵ.

4.2 Logit f-GAN with spline discriminators

For the population analysis in Section 4.1, a discriminator class is assumed to be rich enough to include the nonparametrically optimal discriminator which depends on unknown (ϵ, Q). Because Q can be arbitrary, this nonparametric assumption is inappropriate for sample analysis. Recently, GANs with certain neural network discriminators are shown to achieve sample error bounds matching minimax rates (Gao et al., 2019, 2020). It is interesting to study whether similar results can be obtained when using GANs with simpler and computationally more tractable discriminators.

We propose and study adversarial algorithms, including logit f-GAN in this section and hinge GAN in Section 4.3, each with simple spline discriminators. Define a linear class of pairwise spline functions, denoted as $\mathcal{H}_{\mathrm{sp}}$:

$$h_{\mathrm{sp},\gamma}(x) = \gamma_0 + \gamma_1^{\mathrm{T}}\varphi(x) + \gamma_2^{\mathrm{T}}\mathrm{vec}(\varphi(x) \otimes \varphi(x)),$$

where $\gamma = (\gamma_0, \gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}})^{\mathrm{T}} \in \Gamma$ with $\Gamma = \mathbb{R}^{1+5p+(5p)^2}$ and $\varphi(x) = (\varphi_1^{\mathrm{T}}(x), \ldots, \varphi_5^{\mathrm{T}}(x))^{\mathrm{T}}$. The basis vector $\varphi_l(x) \in \mathbb{R}^p$ is obtained by applying $t \mapsto (t - \xi_l)_+$ componentwise to $x = (x_1, \ldots, x_p)^{\mathrm{T}}$, with the fixed knot $\xi_l = -2, -1, 0, 1,$ or 2 for $l = 1, \ldots, 5$ respectively. For concreteness, assume that every two components of γ_2 are identical if associated with the same product of two components of $\varphi(x)$, that is, γ_2 can be arranged to a symmetric matrix. The preceding specification is sufficient for our theoretical analysis. Nevertheless, similar results can also be obtained, while allowing various changes to the basis functions, for example, adding x as a subvector to $\varphi(x)$. With this change, a function in $\mathcal{H}_{\mathrm{sp}}$ has a main effect term in each x_j, which is a linear spline with fixed knots in $\{-2, -1, 0, 1, 2\}$, and a square or interaction term in each pair (x_{j_1}, x_{j_2}), which is a product of two spline functions in x_{j_1} and x_{j_2} for $1 \le j_1, j_2 \le p$. See Figure 3 for an illustration of the structure of our spline discriminator.

We consider two logit f-GAN methods with an L_1 or L_2 penalty on the discriminator, which lead to meaningful error bounds over the parameter space Θ_1 or Θ_2 respectively under the following conditions on f, in addition to Assumption 1. Among the f-divergences in Table 2, the reverse KL and JS divergences satisfy both Assumptions 1 and 2, and hence the corresponding logit f-GANs achieve sample robust estimation using spline discriminators. The squared Hellinger and reverse χ^2 divergences satisfy Assumption 1, but not the Lipschitz condition in Assumption 2(ii). For such f-divergences, it remains a theoretical question whether sample robust estimation can be achieved using spline discriminators.

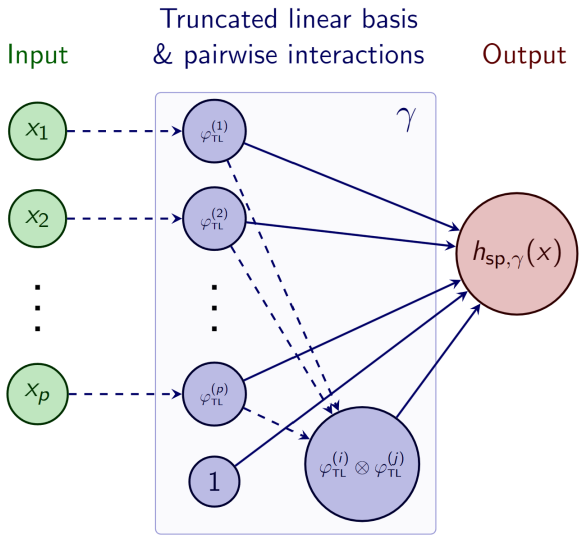

Figure 3: Illustration of our spline discriminator. Dashed lines are fixed transformations with no trainable parameter used and solid lines are linear transformations with parameter γ. For $j = 1, \ldots, p$, $\varphi_{\mathrm{TL}}^{(j)}$ is a vector of truncated linear (TL) basis functions of x_j (the jth element of x) at the fixed knots.

Assumption 2 *Suppose that $f : (0, \infty) \to \mathbb{R}$ is strictly convex and three-times continuously differentiable with $f(1) = 0$ and satisfies the following conditions.*

(i) $f'(\mathrm{e}^u)$ is concave in $u \in \mathbb{R}$.

(ii) $f^{\#}(\mathrm{e}^u)$ is R_1-Lipschitz in $u \in \mathbb{R}$ for a constant $R_1 > 0$.

See Table 2 for validity of conditions (i) and (ii) in various f-divergences, and Remarks 13 and 14 for further discussions.

The first method, L_1 penalized logit f-GAN, is defined by solving

$$\min_{\theta \in \Theta_1} \max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_\theta; h_{\gamma,\mu}) - \lambda_1 \, \mathrm{pen}_1(\gamma) \right\}, \tag{17}$$

where $K_f(P_n, P_\theta; h)$ is $K_f(P_*, P_\theta; h)$ in (1) with P_* replaced by the empirical distribution P_n of $\{X_1, \ldots, X_n\}$, $h_{\gamma,\mu}(x) = h_{\mathrm{sp},\gamma}(x - \mu)$, $\mathrm{pen}_1(\gamma) = \|\gamma_1\|_1 + \|\gamma_2\|_1$, the L_1 norm of γ excluding the intercept γ_0, and $\lambda_1 \geq 0$ is a tuning parameter. In addition to the replacement of P_* by P_n, there are two notable modifications in (17) compared with the population version (1). First, a penalty term is introduced on γ, to achieve suitable control of sampling variation. Second, the discriminator $h_{\gamma,\mu}$ is a spline function with knots depending on μ, the location parameter for the generator. By a change of variables, the non-penalized objective in (17) can be equivalently written as

$$K_f(P_n, P_\theta; h_{\gamma,\mu}) = \mathrm{E}_{P_n - \mu} f'(\mathrm{e}^{h_{\mathrm{sp},\gamma}(x)}) - \mathrm{E}_{P_{0,\Sigma}} f^{\#}(e^{h_{\mathrm{sp},\gamma}(x)}), \tag{18}$$

where $P_n - \mu$ denotes the empirical distribution on $\{X_1 - \mu, \ldots, X_n - \mu\}$. Hence $K_f(P_n, P_\theta; h_{\gamma,\mu})$ is a negative loss for discriminating between the shifted empirical distribution $P_n - \mu$ and the mean-zero generator $P_{0,\Sigma}$. The adaptive choice of knots for the spline discriminator $h_{\gamma,\mu}$ not only is numerically desirable but also facilitates the control of sampling variation in our theoretical analysis. See Propositions 33, 44, 52, and 54. All our sample results such as Theorems 11 and 12 below are non-asymptotic, being valid for any (n, p, ϵ, δ) and Q under the stated conditions.

Theorem 11 *Assume that $\|\Sigma^*\|_{\max} \le M_1$ and f satisfies Assumptions 1–2. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (17). For $\delta < 1/7$, if $\lambda_1 \ge C_1 \left(\sqrt{\log p/n} + \sqrt{\log(1/\delta)/n}\right)$ and $\sqrt{\epsilon} + \sqrt{1/(n\delta)} + \lambda_1 \le C_2$, then with probability at least $1 - 7\delta$ the following bounds hold uniformly over contamination distribution Q,*

$$\|\hat{\mu} - \mu^*\|_\infty \le C\left(\sqrt{\epsilon} + \sqrt{1/(n\delta)} + \lambda_1\right),$$
$$\|\hat{\Sigma} - \Sigma^*\|_{\max} \le C\left(\sqrt{\epsilon} + \sqrt{1/(n\delta)} + \lambda_1\right),$$

where $C_1, C_2, C > 0$ are constants, depending on M_1 and f but independent of (n, p, ϵ, δ).

For L_1 penalized logit f-GAN, Theorem 11 shows that the estimator $(\hat{\mu}, \hat{\Sigma})$ achieves error bounds in the maximum norms in the order $\sqrt{\epsilon} + \sqrt{\log(p)/n}$. These error bounds match sampling errors of order $\sqrt{\log(p)/n}$ in the maximum norms for the standard estimators (i.e., the sample mean and variance) in a multivariate Gaussian model in the case of $\epsilon = 0$. Moreover, up to sampling variation, the error bounds also match the population error bounds of order $\sqrt{\epsilon}$ in the maximum norms with nonparametric discriminators in Theorem 6(i), even though a simple, *linear* class of spline discriminators is used.

The second method, L_2 penalized logit f-GAN, is defined by solving

$$\min_{\theta \in \Theta_2} \max_{\gamma \in \Gamma} \left\{K_f(P_n, P_\theta; h_{\gamma,\mu}) - \lambda_2 \operatorname{pen}_2(\gamma_1) - \lambda_3 \operatorname{pen}_2(\gamma_2)\right\}, \tag{19}$$

where $K_f(P_n, P_\theta; h)$ and $h_{\gamma,\mu}(x)$ are defined as in (17), $\operatorname{pen}_2(\gamma_1) = \|\gamma_1\|_2$ and $\operatorname{pen}_2(\gamma_2) = \|\gamma_2\|_2$, the L_2 norms of γ_1 and γ_2, and $\lambda_2 \ge 0$ and $\lambda_3 \ge 0$ are tuning parameters. Compared with L_1 penalized logit f-GAN (17), the L_2 norms of γ_1 and γ_2 are separately associated with tuning parameters λ_2 and λ_3 in (19), in addition to the change from L_1 to L_2 penalties. As seen from our proofs in Appendices B.2 and B.3, the use of separate tuning parameters λ_2 and λ_3 is crucial for achieving meaningful error bounds in the L_2 and Frobenius norms for simultaneous estimation of (μ^*, Σ^*). Our method does not rely on the use of normalized differences of pairs of the observations to reduce the unknown mean to 0 for scatter estimation as in Diakonikolas et al. (2019).

Theorem 12 *Assume that $\|\Sigma^*\|_{\mathrm{op}} \le M_2$, f satisfies Assumptions 1–2, and $p\epsilon$ is upper bounded by a constant B. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (19). For $\delta < 1/8$, if $\lambda_2 \ge C_1 \left(\sqrt{p/n} + \sqrt{\log(1/\delta)/n}\right)$, $\lambda_3 \ge C_1\sqrt{p}\left(\sqrt{p/n} + \sqrt{\log(1/\delta)/n}\right)$, and $\sqrt{\epsilon} + \sqrt{1/(n\delta)} + \lambda_2 \le C_2$, then with probability at least $1 - 8\delta$ the following bounds hold uniformly over*

contamination distribution Q,

$$\|\hat{\mu} - \mu^*\|_2 \leq C\left(\sqrt{\epsilon} + \sqrt{1/(n\delta)} + \lambda_2\right),$$
$$p^{-1/2}\|\hat{\Sigma} - \Sigma^*\|_{\mathrm{F}} \leq C\left(\sqrt{\epsilon} + \sqrt{1/(n\delta)} + \lambda_2 + \lambda_3/\sqrt{p}\right),$$

where $C_1, C_2, C > 0$ are constants, depending on M_2 and f but independent of (n, p, ϵ, δ) except through the bound B on $p\epsilon$.

For L_2 penalized logit f-GAN, Theorem 12 provides error bounds of order $\sqrt{\epsilon}+\sqrt{p/n}$, in the L_2 and $p^{-1/2}$-Frobenius norms for location and scatter estimation. A technical difference from Theorem 11 is that these bounds are derived under an extraneous condition that $p\epsilon$ is upper bounded. Nevertheless, the error rate, $\sqrt{\epsilon} + \sqrt{p/n}$, matches the population error bounds of order $\sqrt{\epsilon}$ in Theorem 6(i), up to sampling variation of order $\sqrt{p/n}$ in the L_2 and $p^{-1/2}$-Frobenius norms. We defer to Section 4.3 further discussion about the error bounds in Theorems 11–12 compared with minimax error rates.

Remark 13 *There are important implications of Assumption 2(i) together with Assumption 1(ii), based on the fact ("composition rule") that the composition of a non-decreasing concave function and a concave function is concave. First, for convex f, concavity of $f'(\mathrm{e}^u)$ in $u \in \mathbb{R}$ implies Assumption 1(iii), that is, concavity of $f'(t)$ in $t \in (0, \infty)$. This follows by writing $f'(t) = g(\log t)$ and applying the composition rule, where $g(u) = f'(\mathrm{e}^u)$, in addition to being concave, is non-decreasing by convexity of f, and $\log t$ is concave in t. Note that concavity of $f'(t)$ in t may not imply concavity of $f'(\mathrm{e}^u)$ in u, as shown by the Pearson χ^2 in Table 2. Second, for convex and non-increasing f, concavity of $f'(\mathrm{e}^u)$ in $u \in \mathbb{R}$ also implies concavity of $-f^{\#}(\mathrm{e}^u)$ in $u \in \mathbb{R}$. In fact, as mentioned in Remark 3, $f^{\#}(t)$ can be equivalently obtained as $f^{\#}(t) = f^*(f'(t))$, where f^* is the Fenchel conjugate of f (Tan et al., 2019). By the composition rule, $-f^{\#}(\mathrm{e}^u) = g(f'(\mathrm{e}^u))$ is concave, where $g = -f^*$ is concave and non-decreasing by non-increasingness of f.*

Remark 14 *The concavity of $f'(\mathrm{e}^u)$ and $-f^{\#}(\mathrm{e}^u)$ in u from Assumptions 1(ii) and 2(i), as discussed in Remark 13, is instrumental from both theoretical and computational perspectives. These concavity properties are crucial to our proofs of Theorems 11–12 and later Corollary 18(i) in Section 4.4. See Lemmas 31 and 57 in Appendix C. Moreover, the concavity of $f'(\mathrm{e}^u)$ and $-f^{\#}(\mathrm{e}^u)$ in u, in conjunction with the linearity of the spline discriminator $h_{\gamma,\mu}$ in γ, indicates that the objective function $K_f(P_n, P_\theta; h_{\gamma,\mu})$ is concave in γ for any fixed θ. Hence our penalized logit f-GAN (17) or (19) under Assumptions 1–2 can be implemented through nested optimization as in Algorithm 2 with a concave optimizer used to fully update the spline discriminators, as well as through the gradient descent ascent as in Algorithm 1. See Remark 2 for further discussion.*

4.3 Hinge GAN with spline discriminators

We consider two hinge GAN methods with an L_1 or L_2 penalty on the spline discriminator, which leads to theoretically improved error bounds in terms of dependency on (ϵ, p) over the parameter space Θ_1 or Θ_2 respectively, compared with the corresponding logit f-GAN methods in Section 4.2.

The first method, L_1 penalized hinge GAN, is defined by solving

$$\min_{\theta\in\Theta_1}\max_{\gamma\in\Gamma}\left\{K_{\mathrm{HG}}(P_n,P_\theta;h_{\gamma,\mu})-\lambda_1\,\mathrm{pen}_1(\gamma)\right\}, \tag{20}$$

where $K_{\mathrm{HG}}(P_n,P_\theta;h)$ is the hinge objective $K_{\mathrm{HG}}(P_*,P_\theta;h)$ in (3) with P_* replaced by P_n and, similarly as in L_1 penalized logit f-GAN (17), $h_{\gamma,\mu}(x)=h_{\mathrm{sp},\gamma}(x-\mu)$, $\mathrm{pen}_1(\gamma)=\|\gamma_1\|_1+\|\gamma_2\|_1$, and $\lambda_1\geq 0$ is a tuning parameter.

Theorem 15 *Assume that $\|\Sigma^*\|_{\max}\leq M_1$. Let $\hat\theta=(\hat\mu,\hat\Sigma)$ be a solution to (20). For $\delta<1/7$, if $\lambda_1\geq C_1\left(\sqrt{\log p/n}+\sqrt{\log(1/\delta)/n}\right)$ and $\epsilon+\sqrt{\epsilon/(n\delta)}+\lambda_1\leq C_2$, then with probability at least $1-7\delta$ the following bounds hold uniformly over contamination distribution Q,*

$$\|\hat\mu-\mu^*\|_\infty\leq C\left(\epsilon+\sqrt{\epsilon/(n\delta)}+\lambda_1\right),$$
$$\|\hat\Sigma-\Sigma^*\|_{\max}\leq C\left(\epsilon+\sqrt{\epsilon/(n\delta)}+\lambda_1\right),$$

where $C,C_1,C_2>0$ are constants, depending on M_1 but independent of (n,p,ϵ,δ).

For L_1 penalized hinge GAN, Theorem 15 shows that the estimator $(\hat\mu,\hat\Sigma)$ achieves error bounds in the maximum norms in the order $\epsilon+\sqrt{\log(p)/n}$, which improve upon the error rate $\sqrt{\epsilon}+\sqrt{\log(p)/n}$ in terms of dependency on ϵ for L_1 penalized logit f-GAN. This difference can be traced to that in the population error bounds in Theorem 6. Moreover, Theorem 5.1 in Chen et al. (2018) indicates that a minimax lower bound on estimator errors $\|\hat\mu-\mu^*\|_\infty$ or $\|\hat\Sigma-\Sigma^*\|_{\max}$ is also of order $\epsilon+\sqrt{\log(p)/n}$ in Huber's contaminated Gaussian model, where $\sqrt{\log(p)/n}$ is a minimax lower bound in the maximum norms in the case of $\epsilon=0$. Therefore, our L_1 penalized hinge GAN achieves the minimax rates in the maximum norms for Gaussian location and scatter estimation over Θ_1.

The second method, L_2 penalized hinge GAN, is defined by solving

$$\min_{\theta\in\Theta_2}\max_{\gamma\in\Gamma}\left\{K_{\mathrm{HG}}(P_n,P_\theta;h_{\gamma,\mu})-\lambda_2\,\mathrm{pen}_2(\gamma_1)-\lambda_3\,\mathrm{pen}_2(\gamma_2)\right\}, \tag{21}$$

where, similarly as in L_2 penalized logit f-GAN (19), $h_{\gamma,\mu}(x)=h_{\mathrm{sp},\gamma}(x-\mu)$, $\mathrm{pen}_2(\gamma_1)=\|\gamma_1\|_2$ and $\mathrm{pen}_2(\gamma_2)=\|\gamma_2\|_2$, and $\lambda_2\geq 0$ and $\lambda_3\geq 0$ are tuning parameters.

Theorem 16 *Assume that $\|\Sigma^*\|_{\mathrm{op}}\leq M_2$. Let $\hat\theta=(\hat\mu,\hat\Sigma)$ be a solution to (21). For $\delta<1/8$, if $\lambda_2\geq C_1\left(\sqrt{p/n}+\sqrt{\log(1/\delta)/n}\right)$, $\lambda_3\geq C_1\sqrt{p}\left(\sqrt{p/n}+\sqrt{\log(1/\delta)/n}\right)$, and $\sqrt{p}\left(\epsilon+\sqrt{\epsilon/(n\delta)}\right)+\lambda_2\leq C_2$, then with probability at least $1-8\delta$ the following bounds hold uniformly over contamination distribution Q,*

$$\|\hat\mu-\mu^*\|_2\leq C\left(\sqrt{p}\left(\epsilon+\sqrt{\epsilon/(n\delta)}\right)+\lambda_2\right),$$
$$p^{-1/2}\|\hat\Sigma-\Sigma^*\|_{\mathrm{F}}\leq C\left(\sqrt{p}\left(\epsilon+\sqrt{\epsilon/(n\delta)}\right)+\lambda_2+\lambda_3/\sqrt{p}\right),$$

where $C_1,C_2,C>0$ are constants, depending on M_2 but independent of (n,p,ϵ,δ).

For L_2 penalized hinge GAN, Theorem 16 shows that the estimator $(\hat{\mu}, \hat{\Sigma})$ achieves error bounds in the L_2 and $p^{-1/2}$-Frobenius norms in the order $\epsilon\sqrt{p} + \sqrt{p/n}$. On one hand, these error bounds reduce to the same order, $\sqrt{\epsilon} + \sqrt{p/n}$, as those for L_2 penalized logit f-GAN, under the condition that $p\epsilon$ is upper bounded by a constant. On the other hand, when compared with the minimax rates, there remain nontrivial differences between L_2 penalized hinge GAN and logit f-GAN. In fact, the minimax rates in the L_2 and operator norms for location and scatter estimation over Θ_2 is known to be $\epsilon + \sqrt{p/n}$ in Huber's contaminated Gaussian model (Chen et al., 2018). The same minimax rate can also be shown in the $p^{-1/2}$-Frobenius norm for scatter estimation. Then the error rate for L_2 penalized hinge GAN in Theorem 16 matches the minimax rate, and both reduce to the contamination-free error rate $\sqrt{p/n}$, provided that $\epsilon\sqrt{n}$ is bounded by a constant, i.e., $\epsilon = O(\sqrt{1/n})$, independently of p. For L_2 penalized logit f-GAN associated with the reverse KL or JS divergence (satisfying Assumptions 1–2), the error bounds from Theorem 12 match the minimax rate provided both $\epsilon = O(p/n)$ and $\epsilon = O(1/p)$. The latter condition can be restrictive when p is large.

Remark 17 *The two functionals,* $\min(1, h)$ *and* $\min(-1, h)$*, are concave in* h *in the hinge objective* $K_{\mathrm{HG}}(P_n, P_\theta; h)$*. This is reminiscent of the concavity of* $f'(\mathrm{e}^h)$ *and* $-f^{\#}(\mathrm{e}^h)$ *in* h *in the logit* f*-GAN objective* $K_f(P_n, P_\theta; h)$ *under Assumptions 1(ii) and 2(i) as discussed in Remark 14. These concavity properties are crucial to our proofs of Theorems 15–16 and Corollary 18(ii) . See Lemmas 51 and 58 in Appendix C. Moreover, the concavity of* $K_{\mathrm{HG}}(P_n, P_\theta; h)$ *in* h*, together with the linearity of the spline discriminator* $h_{\gamma,\mu}$ *in* γ*, implies that the objective function* $K_{\mathrm{HG}}(P_n, P_\theta; h_{\gamma,\mu})$ *is concave in* γ *for any fixed* θ*. Hence similarly to penalized logit* f*-GAN, our penalized hinge GAN (20) or (21) can also be implemented through nested optimization as in Algorithm 2 with concave inner optimization to update the spline discriminators, as well as the gradient descent ascent as in Algorithm 1.*

4.4 Two-objective GAN with spline discriminators

We study two-objective GANs, where the spline discriminator is trained using the objective function in logit f-GAN or hinge GAN, but the generator is trained using a different objective function.

Consider the following two-objective GAN related to logit f-GANs (17) and (19):

$$\begin{cases} \max\limits_{\gamma\in\Gamma} K_f(P_n, P_\theta; h_{\gamma,\mu}) - \mathrm{pen}(\gamma;\lambda) & \text{with } \theta \text{ fixed,} \\ \min\limits_{\theta\in\Theta} \mathrm{E}_{P_n} f'(\mathrm{e}^{h_{\gamma,\mu}(x)}) - \mathrm{E}_{P_\theta} G(h_{\gamma,\mu}(x)) & \text{with } \gamma \text{ fixed.} \end{cases} \tag{22}$$

Similarly, consider the two-objective GAN related to the hinge GAN (20) and (21):

$$\begin{cases} \max\limits_{\gamma\in\Gamma} K_{\mathrm{HG}}(P_n, P_\theta; h_{\gamma,\mu}) - \mathrm{pen}(\gamma;\lambda) & \text{with } \theta \text{ fixed,} \\ \min\limits_{\theta\in\Theta} \mathrm{E}_{P_n} \min(h_{\gamma,\mu}(x), 1) - \mathrm{E}_{P_\theta} G(h_{\gamma,\mu}(x)) & \text{with } \gamma \text{ fixed.} \end{cases} \tag{23}$$

Here $\mathrm{pen}(\gamma;\lambda)$ is an L_1 penalty, $\lambda_1(\|\gamma_1\|_1 + \|\gamma_2\|_1)$ and Θ is $\Theta_1 = \{(\mu, \Sigma) : \mu \in \mathbb{R}^p, \|\Sigma\|_{\max} \le M_1\}$ as in (17), or $\mathrm{pen}(\gamma;\lambda)$ is an L_2 penalty $\lambda_2\|\gamma_1\|_2 + \lambda_3\|\gamma_2\|_2$ and Θ is $\Theta_2 = \{(\mu, \Sigma) : \mu \in \mathbb{R}^p, \|\Sigma\|_{\mathrm{op}} \le M_2\}$ as in (19), and G is a function satisfying Assumption 3. Note that the discriminator $h_{\gamma,\mu}$ is a spline function with knots depending on μ, so that $\mathrm{E}_{P_n} f'(\mathrm{e}^{h_{\gamma,\mu}(x)})$

cannot be dropped in the optimization over θ in (22) or (23). We show that the two-objective logit f-GAN and hinge GAN achieve similar error bounds as the corresponding one-objective versions in Theorems 11–16.

Assumption 3 *Function G in (22) or (23) is convex and strictly increasing. Hence the inverse function G^{-1} exists and is concave and strictly increasing.*

Corollary 18 *(i) If $\hat{\theta}$ is replaced by a solution to the alternating optimization problem (22) with the L_1 or L_2 penalty on γ as in (17) or (19) and the corresponding choice of Θ, then the results in Theorem 11 or 12 remains valid respectively.*

(ii) If $\hat{\theta}$ is replaced by a solution to the alternating optimization problem (23) with the L_1 or L_2 penalty on γ as in (20) or (21) and the corresponding choice of Θ, then the results in Theorem 15 or 16 remains valid respectively.

The two-objective GANs studied in Corollary 18 differ slightly from existing ones as described in (5)–(7), due to the use of the discriminator $h_{\gamma,\mu}$ depending on μ to facilitate theoretical analysis as mentioned in Section 4.2. If $h_{\gamma,\mu}$ were replaced by a discriminator h_γ defined independently of θ, then taking $K_f = K_{\mathrm{JS}}$ and $G(h) = -\log(1+\mathrm{e}^{-h})$ or $G(h) = -h$ in (22) reduces to GAN with log$D$ trick (5) or calibrated rKL-GAN (6) respectively, and taking $K_f = K_{\mathrm{HG}}$ and $G(h) = -h$ in (23) reduces to geometric GAN (7).

5. Discussion

5.1 GANs with data transformation

Compared with the usual formulations (1) and (3), our logit f-GAN and hinge GAN methods in Sections 4.2–4.3 involve a notable modification that both the real and fake data are discriminated against each other after being shifted by the current location parameter. Without the modification, a direct approach based on logit f-GAN would use the objective function

$$K_f(P_n, P_\theta; h_{\mathrm{sp},\gamma}) = \mathrm{E}_{P_n} f'(\mathrm{e}^{h_{\mathrm{sp},\gamma}(x)}) - \mathrm{E}_{P_{\mu,\Sigma}} f^{\#}(e^{h_{\mathrm{sp},\gamma}(x)}), \tag{24}$$

where the real data and the Gaussian fake data generated from standard noises are discriminated again each other given the parameters (μ, Σ). The idea behind our modification can be extended by allowing both location and scatter transformation. For example, consider logit f-GAN with full transformation:

$$\min_{\theta\in\Theta} \max_{\gamma\in\Gamma} \left\{K_f(P_n, P_\theta; h_{\gamma,\mu,\Sigma}) - \mathrm{pen}(\gamma;\lambda)\right\}, \tag{25}$$

where K_f is the logit f-GAN objective as in (17) and (19), $h_{\gamma,\mu,\Sigma}(x) = h_{\mathrm{sp},\gamma}(\Sigma^{-1/2}(x-\mu))$ and $\mathrm{pen}(\gamma;\lambda)$ is an L_1 or L_2 penalty term. The discriminator $h_{\gamma,\mu,\Sigma}(x)$ is obtained by applying $h_{\mathrm{sp},\gamma}(\cdot)$ with fixed knots to the transformed data $\Sigma^{-1/2}(x-\mu)$. Similarly to (18), the non-penalized objective in (25) can be equivalently written as

$$K_f(P_n, P_\theta; h_{\gamma,\mu,\Sigma}) = \mathrm{E}_{\Sigma^{-1/2}(P_n-\mu)} f'(\mathrm{e}^{h_{\mathrm{sp},\gamma}(x)}) - \mathrm{E}_{P_{0,I}} f^{\#}(e^{h_{\mathrm{sp},\gamma}(x)}), \tag{26}$$

where $\Sigma^{-1/2}(P_n - \mu)$ denotes the empirical distribution on $\{\Sigma^{-1/2}(X_1 - \mu), \ldots, \Sigma^{-1/2}(X_n - \mu)\}$. Compared with (18) and (24), there are two advantages of using (26) with full transformation. First, due to both location and scatter transformation, logit f-GAN (25), but not (17) or (19), can be shown to be affine equivariant. Second, the transformed real data and the standard Gaussian noises in (26) are discriminated against each other given the current parameters (μ, Σ), while employing the spline discriminators $h_{\mathrm{sp},\gamma}(x)$ with knots fixed at $\{-2, -1, 0, 1, 2\}$. Because standard Gaussian data are well covered by the grid formed from these marginal knots, the discrimination involved in (26) can be informative even when the parameters (μ, Σ) are updated. The discrimination involved in (24) may be problematic when employing the fixed-knot spline discriminators, because both the real and fake data may not be adequately covered by the grid formed from the knots.

From the preceding discussion, it can be more desirable to incorporate both location and scatter transformation as in (26) than just location transformation as in (18), which only aligns the centers, but not the scales and correlations, of the Gaussian fake data with the knots in the spline discriminators. As mentioned in Section 4.2, our sample analysis exploits the location transformation in establishing certain concentration properties in the proofs. On the other hand, our current proofs are not directly applicable while allowing both location and scatter transformation. It is desired in future work to extend our theoretical analysis in this direction.

5.2 Comparison with Gao et al. (2020)

We first point out a connection between logit f-GANs and the GANs based on proper scoring rules in Gao et al. (2020). For a convex function $g : (0, 1) \to \mathbb{R}$, a proper scoring rule can be defined as (Savage, 1971; Buja et al., 2005; Gneiting and Raftery, 2007)

$$S_g(\eta, 1) = g(\eta) + (1 - \eta)g'(\eta), \quad S_g(\eta, 0) = g(\eta) - \eta g'(\eta).$$

The population verion of the GAN studied in Gao et al. (2020) is defined as

$$\min_{\theta \in \Theta} \max_{\gamma \in \Gamma} L_g(P_*, P_\theta; q_\gamma), \tag{27}$$

where $q_\gamma(x) \in [0, 1]$, also called a discriminator, represents the probability that an observation x comes from P_* rather than P_θ, and

$$L_g(P_*, P_\theta; q) = (1/2)\left\{\mathrm{E}_{P_*} S_g(q(x), 1) + \mathrm{E}_{P_\theta} S_g(q(x), 0)\right\} - g(1/2).$$

The objective $L_g(P_*, P_\theta; q)$ is shown to be a lower bound, being tight if $q = 2\,\mathrm{d}P_*/\mathrm{d}(P_* + P_\theta)$, for the divergence $D_{g_0}(P_* \| (P_* + P_\theta)/2)$, where $g_0(t) = g(t/2) - g(1/2)$ for $t \in (0, 2)$. For example, taking $g(\eta) = \eta \log \eta + (1 - \eta) \log(1 - \eta)$ leads to the log score, $S_g(\eta, 1) = \log \eta$ and $S_g(\eta, 0) = \log(1 - \eta)$. The corresponding objective function $L_g(P_*, P_\theta; q_\gamma)$ reduces to the expected log-likelihood with discrimination probability $q_\gamma(x)$ as used in Goodfellow et al. (2014). We show that if $q_\gamma(x)$ is specified as a sigmoid probability, then $L_g(P_*, P_\theta; q_\gamma)$ can be equivalently obtained as a logit f-GAN objective for a suitable choice of f.

Proposition 19 *Suppose that the discriminator is specified as* $q_\gamma(x) = \mathrm{sigmoid}(h_\gamma(x))$. *Then* $L_g(P_*, P_\theta; q_\gamma) = K_f(P_*, P_\theta; h_\gamma)$ *for* K_f *defined in (1) and* $f(t) = \frac{1+t}{2} g_0(\frac{2t}{1+t})$ *satisfying that* $D_{g_0}(P_* \| (P_* + P_\theta)/2) = D_f(P_* \| P_\theta)$.

In contrast with $h_\gamma(x)$ parameterized as a pairwise spline function, Gao et al. (2020) studied robust estimation in Huber's contaminated Gaussian model, where $q_\gamma(x)$ is parameterized as a neural network with two or more layers and sigmoid activations in the top and bottom layers. In the case of two layers, the neural network in Gao et al. (2020), Section 4, is defined as

$$q_\gamma(x) = \text{sigmoid}(h_\gamma(x)), \quad h_\gamma(x) = \sum_{j=1}^{J} \gamma_j^{(1)} \text{sigmoid}(\gamma_j^{(2)\text{T}} x + \gamma_{0j}^{(2)}), \tag{28}$$

where $(\gamma_j^{(2)}, \gamma_{j0}^{(2)})$, $j = 1, \ldots, J$, are the weights and intercepts in the bottom layer, and $\gamma_j^{(1)}$, $j = 1, \ldots, J$, are the weights in the top layer constrained such that $\sum_{j=1}^{J} |\gamma_j^{(1)}| \le \kappa$ for a tuning parameter κ. Assume that $g(\eta)$ is three-times continuously differentiable at $\eta = 1/2$, $g''(1/2) > 0$, and for a universal constant $c_0 > 0$,

$$2g''(1/2) \ge g'''(1/2) + c_0, \tag{29}$$

Then Gao et al. (2020) showed that the location and scatter estimators from the sample version of (27) with discriminator (28) achieve the minimax error rates, $O(\epsilon + \sqrt{p/n})$, in the L_2 and operator norms, provided that $\kappa = O(\epsilon + \sqrt{p/n})$ among other conditions. However, with sigmoid activations used inside $h_\gamma(x)$, the sample objective $L_g(P_n, P_\theta; q_\gamma)$ may exhibit a complex, non-concave landscape in γ, which makes minimax optimization difficult.

There is also a subtle issue in how the above result from Gao et al. (2020) can be compared with even our population analysis for minimum f-divergence estimation, i.e., population versions of GANs with nonparametric discriminators. In fact, condition (29) can be directly shown to be equivalent to saying that $\frac{\mathrm{d}^2}{\mathrm{d}u^2} f'(\mathrm{e}^u)|_{u=0} \ge c_0$ for f associated with g in Proposition 19. This condition can be satisfied, while Assumption 1 is violated, for example, by the choice $g(\eta) = (\eta - 1)\log(\eta/(2-\eta))$ and $f(t) = \{(t-1)\log t\}/2$, corresponding to the mixed KL divergence $D_{\text{KL}}(P||Q)/2 + D_{\text{KL}}(Q||P)/2$. As shown in Figure 2, minimization of the mixed KL does not in general lead to robust estimation. Hence it seems paradoxical that minimax error rates can be achieved by the GAN in Gao et al. (2020) with its objective function derived from the mixed KL. On the other hand, a possible explanation can be seen as follows. By the sigmoid activation and the constraint $\sum_{j=1}^{J} |\gamma_j^{(1)}| \le \kappa$, the log-odds discriminator $h_\gamma(x)$ in (28) is forced to be bounded, $|h_\gamma(x)| \le \kappa$, where κ is further assumed to small, of the same order as the minimax rate $O(\epsilon + \sqrt{p/n})$. As a result, maximization of the population objective $L_g(P_*, P_\theta; q_\gamma)$ over such constrained discriminators may produce a divergence with a substantial gap to the actual divergence $D_f(P_* \| P_\theta)$ for any fixed θ. Instead, the implied divergence measure may behave more similarly as the total variation $D_{\text{TV}}(P_* \| P_\theta)$ than as $D_f(P_* \| P_\theta)$, due to the boundedness of $h_\gamma(x)$ by a sufficiently small κ, so that minimax error rates can still be achieved.

6. Simulation studies

We conducted simulation studies to compare the performance of our logit f-GAN and hinge GAN methods with several existing methods in various settings depending on Q, ϵ, n, and p. Results about error dependency on ϵ are provided in Section 6.3 and those about dependency on n and p are presented in Appendix A. Two contamination distributions Q are

considered to allow different types of contaminations. In the arXiv preprint of the paper (Wang and Tan, 2021), only low-dimensional settings are studied, with p between 5–20 and n between 500–4000. In the current paper, relatively high-dimensional settings are studied, with p between 25–100 and n between 5000–50000. In such settings, the previous implementation of our methods based on nested optimization (Algorithm 2) becomes computationally costly, and hence the current implementation of our methods follows the style of alternating gradient updates in Algorithm 1, but with Adam used (Kingma and Ba, 2015) instead of vanilla gradient updates. As discussed in Remark 2, the concavity of our GAN objectives in the discriminators makes it possible to treat local minimax points as a generally valid surrogate for global solutions. In addition, training of our methods also benefits from the fact that the discriminators can be updated without ever being trapped in local maxima and hence the generators can be consistently pushed into the right direction.

6.1 Implementation of methods

Our methods can be implemented in the style of either nested optimization (Algorithm 2) or alternating gradient updates (Algorithm 1). Source code for our methods is available at `https://github.com/LMC4S/robust-spline-GAN` for nested optimization and `https://github.com/LMC4S/robust-spline-GAN-pytorch` for alternating gradient updates.

We refer to our arXiv preprint (Wang and Tan, 2021) for the former implementation which is suitable in low-dimensional settings and present only the latter implementation which is more cost-effective in relatively high-dimensional settings. Our detailed pseudo code is shown as Algorithm 3 in Appendix A, including the initial values (μ_0, Σ_0) and the learning rates. The penalized GAN objective function $K(\theta, \gamma; \lambda)$ is defined as in (25) for logit f-GAN or with K_f replaced by K_{HG} for hinge GAN. As discussed in Section 5.1, this scheme allows adequate discrimination between the back-transformed real data, $\Sigma^{-1/2}(x - \mu)$, and the standard Gaussian noises using spline discriminators with *fixed knots*. Below we briefly discuss the alternating gradient updates and penalty choices.

With spline discriminators, the training objective $K(\theta, \gamma; \lambda)$ is concave in the discriminator parameter γ and hence the discriminator can be consistently updated to provide a proper updating direction for the generator. Instead of vanilla gradient updates, we use Adam (Kingma and Ba, 2015) with a momentum and an adaptive learning rate to alternately update both the discriminator and the generator in the style of Algorithm 1. The introduction of the momentum helps to overcome possible local minima for the generator and also accelerates the training for the discriminator.

As dictated by our theory, we employ L_1 or L_2 penalties on the spline discriminators to control sampling variation, especially when the sample size n is relatively smaller compared to the dimension of the discriminator parameter γ. Numerically, these penalties help stabilize the training process by restricting the discriminator power in the early stage. We tested our methods under different penalty levels and identified default choices of λ for our rKL and JS logit f-GANs and hinge GAN. These penalty choices are then fixed in all our subsequent simulations. See Appendix A.2 for results from our tuning experiments.

For comparison, we also implement 5 existing methods for robust estimation.

- *JS-GAN* (Gao et al., 2020). We use the code from Gao et al. (2020) with minimal modification. The batch size is set to 1/10 of the data size because the default choice

500 is too large in our experiment settings. We use the network structure p-$2p$-$\lfloor p/2 \rfloor$-1 with LeakyReLU and Sigmoid activations as recommended in Gao et al. (2020).

- *Kendall's τ with MAD* (Loh and Tan, 2018). Kendall's τ (Kendall, 1938) is used to estimate the correlations after sine transformation and the median absolute deviation (MAD) (Hampel, 1974) is used to estimate the scales. We use `stats.kendalltau` and `stats.median_abs_deviation` from Python module `SciPy` to compute Kendall's τ correlations and MAD scale (`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html`).

- *Spearman's ρ with Q_n-estimator* (Öllerer and Croux, 2015). The Q_n-estimator (Rousseeuw and Croux, 1993) is used for scale estimation and Spearman's ρ (Spearman, 1987) is used with sine transformation for correlation estimation. We use the R function `corollary` to compute Spearman's ρ correlations and the `Qn` function in R package `robustbase` (`https://cran.r-project.org/web/packages/robustbase`).

- *MCD* (Rousseeuw, 1985). The minimum covariance determinant (MCD) estimator is a high-breakdown robust method and is shown to be superior to the Minimum volume ellipsoid (MVE) estimator in statistical efficiency (Butler et al., 1993). We use `covariance.MinCovDet` in Python module `scikit-learn` for implementation. (`https://scikit-learn.org/stable/modules/generated/sklearn.covariance.MinCovDet.html`)

- *Tyler's M-estimator* (Tyler, 1987). This method is included for completeness, being designed for multivariate scatter estimation from elliptical distributions, not Huber's contaminated Gaussian distribution. The estimated scatter matrix is uniquely defined subject to the constraint that the determinant is 1. To facilitate comparison in terms of variance matrix estimation, we rescale the scatter matrix such that its determinant matches that of the true variance matrix Σ^*, even though this may lead to some unfair advantage. We use R package `fastM` for implementation (`https://cran.r-project.org/web/packages/fastM`).

In our experiments, we focus on comparing the performance of existing and proposed methods in terms of scatter estimation (i.e., variance matrix estimation). Tyler's M-estimator, Kendall's τ with MAD, and Spearman's ρ with Q_n deal with scatter estimation only and hence the locations are set to the true means as mentioned in Section 2. The other methods handle both location and scatter estimation.

There are also robust S- and M-estimators, for example, based on translated or Tukey's biweight functions, which are shown to achieve a high-breakdown property (Rousseeuw, 1985; Rocke, 1996). Several such estimators were included in our previous low-dimensional experiments (Wang and Tan, 2021). However, the existing R packages for those methods fail to run successfully in our relatively high-dimensional settings and hence those methods are not considered in the current experiments.

6.2 Simulation settings

The uncontaminated distribution is $\mathrm{N}(0, \Sigma^*)$ where Σ^* is a Toeplitz matrix with (i, j) component equal to $(1/2)^{|i-j|}$. The location parameter is unknown and estimated together with

the variance matrix, except for Tyler's M-estimator, Kendall's τ, and Spearman's ρ. Consider two contamination distributions Q of different types. Denote a $p \times p$ identity matrix as I_p and a p-dimensional vector of ones as 1_p.

- $Q = \text{Cauchy}\left(2.25c, \frac{1}{3}I_p\right)$ where $c = (1, -1, 1, -1, 1, \dots)$ is a p-dimensional vector of alternating ± 1. In this setting, the majority of contaminated points may not be seen as outliers marginally in each coordinate. On the other hand, these contaminated points can be easily separated as outliers from the uncontaminated Gaussian distribution in higher dimensions.

- $Q = \text{Cauchy}(51_p, 5I_p)$. Contaminated points may lie in both low-density and high-density regions of the uncontaminated Gaussian distribution. The majority of contaminated points are outliers that are far from the uncontaminated data, and there are also contaminated points that are enclosed by the uncontaminated points.

The Cauchy contamination, although being extreme, is chosen to assess our theoretical results, which are uniform over all possible contaminations. Compared with Gaussian contamination distributions, the setting also makes training of GANs more difficult because Cauchy does not have any finite moments and some data points can be excessive outliers. The success of our methods in the presence of Cauchy contamination, as shown below, provides a strong support for our methods in handling all possible contaminations. See Wang and Tan (2021) for numerical studies with Gaussian contaminations in low-dimensional settings, where similar patterns are observed as reported here.

6.3 Experiment results

Table 3 summarizes scatter estimation errors in the maximum norm from L_1 penalized hinge GAN and logit f-GANs and existing methods, where $p = 100$, $n = 20000$, and ϵ increases from 0% to 20%. See Appendix A.3 for additional results about error dependency on n and p. The errors are obtained by averaging 20 repeated runs and the numbers in brackets are standard deviations. The JS logit f-GAN has the best performance, followed closely by rKL logit f-GAN and hinge GAN and then with more noticeable differences by JS-GAN in Gao et al. (2020). The MCD performs among the best when there is no contamination ($\epsilon = 0$), but its performance deteriorates considerably as ϵ increases to 20%, especially with the first contamination. The pairwise methods, Kendall's τ with MAD and Spearman's ρ with Q_n-estimator, have poor performances as expected from Figure 1. Estimation errors in the Frobenius norm from our L_2 penalized GAN methods and existing methods are shown in Table 4. We observe a similar pattern of comparison as in Table 3, except that the hinge GAN achieves a slight lead.

From Tables 3–4, we see that the estimation errors of our GAN methods, as well as other methods, increase as ϵ increases. However, the dependency on ϵ is not precisely linear for the hinge GAN, and not in the order $\sqrt{\epsilon}$ for the two logit f-GANs. This does not violate our theoretical bounds, which are derived to hold over all possible contamination distributions, i.e., for the worst scenario of contamination. For specific contamination settings, it is possible for logit f-GAN to outperform hinge GAN, and for each method to achieve a better error dependency on ϵ than in the worst scenario. For further understanding, we present in Figure 8 (Appendix A.5) a comparison between two types of contamination

ϵ (%)	hinge GAN	JS logit f-GAN	rKL logit f-GAN	GYZ JS-GAN	Kendall_MAD	Spearman_Qn	MCD	Tyler_M
0	0.0299 (0.0027)	0.0304 (0.0027)	0.0321 (0.0042)	0.0360 (0.006)	0.0445 (0.0057)	0.0385 (0.0032)	**0.0296 (0.0022)**	0.0299 (0.0025)
				$Q \sim \text{Cauchy}\left(2.25c, \frac{1}{3}I_p\right)$				
5	0.0333 (0.0031)	**0.0302 (0.0027)**	0.0303 (0.0028)	0.0472 (0.0116)	0.1651 (0.0058)	0.1989 (0.0051)	0.0470 (0.0264)	0.3138 (0.0142)
10	0.0356 (0.0032)	**0.0309 (0.0036)**	0.0311 (0.0025)	0.0482 (0.0123)	0.3165 (0.0063)	0.3906 (0.0070)	0.3115 (0.0089)	0.7810 (0.0190)
20	0.0394 (0.0047)	**0.0341 (0.0033)**	0.0343 (0.0034)	0.0527 (0.0096)	0.7514 (0.0122)	0.8297 (0.0055)	0.6510 (0.0065)	1.8045 (0.0368)
				$Q \sim \text{Cauchy}(51_p, 5I_p)$				
5	0.0354 (0.0038)	**0.0305 (0.0025)**	0.0340 (0.0034)	0.0416 (0.0054)	0.1703 (0.0068)	0.2451 (0.0058)	0.0410 (0.0048)	0.1228 (0.0047)
10	0.0374 (0.0036)	**0.0319 (0.0035)**	0.0361 (0.0041)	0.0450 (0.0085)	0.3287 (0.0081)	0.5167 (0.0072)	0.0540 (0.0044)	0.2645 (0.0062)
20	0.0433 (0.0045)	**0.0349 (0.0033)**	0.0385 (0.0043)	0.0483 (0.0113)	0.8071 (0.0127)	1.3132 (0.0104)	0.0821 (0.0040)	0.5850 (0.0118)

Table 3: Comparison of existing methods and proposed L_1 penalized GAN methods ($p = 100, n = 20000$, and varying ϵ from 0% to 20%). Estimation error of the variance matrix is reported in the maximum norm $\|\cdot\|_{\max}$.

ϵ (%)	hinge GAN	JS logit f-GAN	rKL logit f-GAN	GYZ JS-GAN	Kendall_MAD	Spearman_Qn	MCD	Tyler_M
0	0.7385 (0.0136)	0.7395 (0.0117)	0.7413 (0.0127)	0.8309 (0.0217)	0.766 (0.0139)	0.7793 (0.0189)	**0.7333 (0.0124)**	0.7357 (0.0121)
				$Q \sim \text{Cauchy}\left(2.25c, \frac{1}{3}I_p\right)$				
5	**0.7581 (0.0117)**	0.7600 (0.0116)	0.7612 (0.0122)	0.9858 (0.0533)	11.3637 (0.1097)	11.4136 (0.1443)	1.3872 (2.7002)	25.4480 (1.2743)
10	**0.7781 (0.0123)**	0.7822 (0.01)	0.7831 (0.0107)	1.0183 (0.0783)	24.2221 (0.2151)	24.9319 (0.2416)	26.8421 (0.6546)	68.9657 (1.7968)
20	**0.8230 (0.0127)**	0.8257 (0.0108)	0.8324 (0.0139)	1.0726 (0.0509)	56.9405 (0.4234)	57.8273 (0.3415)	57.2124 (0.7063)	165.0762 (3.6783)
				$Q \sim \text{Cauchy}(51_p, 5I_p)$				
5	**0.7599 (0.0115)**	0.7605 (0.0104)	0.7706 (0.0105)	0.9423 (0.0274)	10.2662 (0.1475)	10.6577 (0.1949)	0.7827 (0.0153)	9.6110 (0.4390)
10	**0.7804 (0.0103)**	0.7814 (0.0112)	0.7897 (0.0096)	0.9926 (0.0394)	21.9943 (0.2992)	24.4219 (0.3122)	0.8598 (0.0165)	23.4787 (0.5847)
20	**0.8252 (0.0114)**	0.8261 (0.0129)	0.8376 (0.0125)	1.0623 (0.0627)	52.9465 (0.5247)	65.1641 (0.5581)	1.0715 (0.0214)	54.9783 (1.2116)

Table 4: Comparison of existing methods and proposed L_2 penalized GAN methods ($p = 100, n = 20000$, and varying ϵ from 0% to 20%). Estimation error of the variance matrix is reported in the Frobenius norm $\|\cdot\|_{\mathrm{F}}$.

settings for GANs at the population level, similarly to Figure 2. One type may represent the worst-case contamination in terms of dependency on ϵ, and the other type is based on the second contamination studied.

7. Main proofs

We present main proofs of Theorems 6 and 15 in this section. The main proofs of the other results and details of all main proofs are provided in Appendices B and C.

At the center of our proofs is a unified strategy designed to establish error bounds for GANs. See, for example, the two-sided bounds of the penalized GAN objective with optimized discriminator in (31) and (37). To derive the upper bounds, we apply the robustness property of TV or f-divergence under Assumptions 1–2 to remove the impact of contamination, and then develop suitable concentration properties based on Gaussian or sub-Gaussian while leveraging the concavity in updating the spline discriminators for hinge GAN or logit f-GAN (as discussed in Remarks 14 and 17). These can be seen from the proofs of Propositions 33, 44, 52, and 54 in Appendix C. To derive the lower bounds, we exploit the fact that it is sufficient to consider a subclass of bounded ramp functions constructed from unbounded spline functions, and then develop desirable concentration properties over the ramp or product ramp functions under a general contaminated distribution. These can be seen from the proofs of Propositions 37, 47, 53, and 55 in Appendix C. Finally, we deduce estimation error bounds by showing that the expectations of ramp or product ramp func-

tions are locally linear in the location, scale, and correlation of the underlying Gaussian distribution; see Lemmas 38 and 40 in Appendix C, with a novel application of Stein's lemma.

7.1 Proof of Theorem 6

We state and prove the following result which implies Theorem 6.

Proposition 20 *Let $\Theta_0 = \{(\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma$ is a $p \times p$ variance matrix $\}$.*

(i) Assume that f satisfies Assumption 1, and $\epsilon \in [0, \epsilon_0]$ for a constant $\epsilon_0 \in [0, 1/2)$. Let $\bar{\theta} = \mathrm{argmin}_{\theta \in \Theta_0} D_f(P_\epsilon || P_\theta)$. If $\mathrm{Err}_{f0}(\epsilon) \leq a$ for a constant $a \in [0, 1/2)$, then we have

$$\|\bar{\mu} - \mu^*\|_2 \leq S_{1,a} \|\Sigma^*\|_{\mathrm{op}}^{1/2} \mathrm{Err}_{f0}(\epsilon),$$
$$\|\bar{\mu} - \mu^*\|_\infty \leq S_{1,a} \|\Sigma^*\|_{\max}^{1/2} \mathrm{Err}_{f0}(\epsilon),$$

where $S_{1,a} = \{\Phi'(\Phi^{-1}(1/2 + a))\}^{-1}$ and $\mathrm{Err}_{f0}(\epsilon) = \sqrt{-2(f''(1))^{-1} f'(1 - \epsilon_0)\epsilon} + \epsilon$. If further $\mathrm{Err}_{f0}(\epsilon) \leq a/(1 + S_{1,a})$, then

$$\|\bar{\Sigma} - \Sigma^*\|_{\mathrm{op}} \leq 2S_{3,a} \|\Sigma^*\|_{\mathrm{op}} \mathrm{Err}_{f0}(\epsilon) + S_{3,a}^2 \|\Sigma^*\|_{\mathrm{op}} (\mathrm{Err}_{f0}(\epsilon))^2, \tag{30}$$
$$\|\bar{\Sigma} - \Sigma^*\|_{\max} \leq 4S_{3,a} \|\Sigma^*\|_{\max} \mathrm{Err}_{f0}(\epsilon) + 2S_{3,a}^2 \|\Sigma^*\|_{\max} (\mathrm{Err}_{f0}(\epsilon))^2,$$

where $S_{3,a} = S_{2,a}(1 + S_{1,a})$, $S_{2,a} = \{\sqrt{z_0/2}\, \mathrm{erf}'(\sqrt{2/z_0}\, \mathrm{erf}^{-1}(1/2 + a))\}^{-1}$, and the constant z_0 is defined such that $\mathrm{erf}(\sqrt{z_0/2}) = 1/2$. The same inequality as (30) also holds with $\|\bar{\Sigma} - \Sigma^*\|_{\mathrm{op}}$ replaced by $p^{-1/2} \|\bar{\Sigma} - \Sigma^*\|_{\mathrm{F}}$.

(ii) Let $\bar{\theta} = \mathrm{argmin}_{\theta \in \Theta_0} D_{\mathrm{TV}}(P_\epsilon || P_\theta)$. Then the statements in (i) hold with $\mathrm{Err}_{f0}(\epsilon)$ replaced by $\mathrm{Err}_{h0}(\epsilon) = 2\epsilon$ throughout.

Proof [Proof of Proposition 20] (i) Our main strategy is to show the following inequalities hold:

$$d(\bar{\theta}, \theta^*) - \Delta_1(\epsilon) \leq \sqrt{D_f(P_\epsilon || P_{\bar{\theta}})} \leq \Delta_2(\epsilon, f), \tag{31}$$

where $\Delta_1(\epsilon)$ and $\Delta_2(\epsilon, f)$ are bias terms, depending on ϵ and (ϵ, f) respectively and $d(\bar{\theta}, \theta^*)$ is the total variation $D_{\mathrm{TV}}(P_{\bar{\theta}} \| P_{\theta^*})$ or simply $\mathrm{TV}(P_{\bar{\theta}}, P_{\theta^*})$. Under certain conditions, $d(\bar{\theta}, \theta^*)$ delivers upper bounds, up to scaling constants, on the estimation bias to be controlled, $\|\bar{\mu} - \mu^*\|_\infty$, $\|\bar{\mu} - \mu^*\|_2$, $\|\bar{\Sigma} - \Sigma^*\|_{\max}$, and $\|\bar{\Sigma} - \Sigma^*\|_{\mathrm{op}}$.

(Step 1) The upper bound in (31) follows from Lemma 31 (iv): for any f satisfying Assumption 1 and any $\epsilon \in [0, \epsilon_0]$, we have

$$D_f(P_\epsilon || P_{\bar{\theta}}) \leq D_f(P_\epsilon || P_{\theta^*}) \leq -f'(1 - \epsilon_0)\epsilon = \Delta_2^2(\epsilon, f),$$

where $\Delta_2(\epsilon, f) = \sqrt{-f'(1 - \epsilon_0)\epsilon}$. The constant $-f'(1 - \epsilon_0)$ is nonnegative because f is non-increasing by Assumption 1 (ii).

(Step 2) We show the lower bound in (31) as follows:

$$d(\bar{\theta}, \theta^*) \leq \mathrm{TV}(P_{\bar{\theta}}, P_\epsilon) + \mathrm{TV}(P_\epsilon, P_{\theta^*}) \leq \mathrm{TV}(P_{\bar{\theta}}, P_\epsilon) + \Delta_1(\epsilon) \tag{32}$$
$$\leq \sqrt{2(f''(1))^{-1} D_f(P_\epsilon || P_{\bar{\theta}})} + \Delta_1(\epsilon), \tag{33}$$

where $\Delta_1(\epsilon) = \epsilon$. Line (32) follows by the triangle inequality and the fact that $\mathrm{TV}(P_\epsilon, P_{\theta^*}) \leq \epsilon \mathrm{TV}(P_Q, P_{\theta^*}) \leq \epsilon$. Line (33) follows from Lemma 27: for any f-divergence satisfying Assumption 1 (iii), we have

$$D_f(P_\epsilon || P_{\bar{\theta}}) \geq \frac{f''(1)}{2} \mathrm{TV}(P_\epsilon, P_{\bar{\theta}})^2.$$

The scaling constant, $\inf_{t \in (0,1]} f''(t)/2$, in Lemma 27 reduces to $f''(1)/2$, because f'' is non-increasing by Assumption 1 (iii).

(Step 3) Combining the lower and upper bounds in (31), we have

$$d(\bar{\theta}, \theta^*) \leq \sqrt{2(f''(1))^{-1}} \Delta_2(\epsilon, f) + \Delta_1(\epsilon) = \mathrm{Err}_{f0}(\epsilon),$$

where $\mathrm{Err}_{f0}(\epsilon) = \sqrt{-2(f''(1))^{-1} f'(1-\epsilon_0)\epsilon} + \epsilon$. The location result then follows from Proposition 29 provided that $\mathrm{Err}_{f0}(\epsilon) \leq a$ for a constant $a \in [0, 1/2)$. The variance matrix result follows if $\mathrm{Err}_{f0}(\epsilon) \leq a/(1 + S_{1,a})$.

(ii) For the TV minimizer $\bar{\theta}$, Steps 1 and 2 in (i) can be combined to directly obtain an upper bound on $d(\bar{\theta}, \theta^*)$ as follows:

$$\begin{aligned} d(\bar{\theta}, \theta^*) &\leq \mathrm{TV}(P_{\bar{\theta}}, P_\epsilon) + \mathrm{TV}(P_\epsilon, P_{\theta^*}) && (34) \\ &\leq 2\mathrm{TV}(P_\epsilon, P_{\theta^*}) && (35) \\ &\leq 2\epsilon. && (36) \end{aligned}$$

Line (34) is due to the triangle inequality. Line (35) follows because $\mathrm{TV}(P_{\bar{\theta}}, P_\epsilon) \leq \mathrm{TV}(P_{\theta^*}, P_\epsilon) = \mathrm{TV}(P_\epsilon, P_{\theta^*})$ by the definition of $\bar{\theta}$ and the symmetry of TV. Line (36) follows because $\mathrm{TV}(P_\epsilon, P_{\theta^*}) \leq \epsilon \mathrm{TV}(P_Q, P_{\theta^*}) \leq \epsilon$ as in (32).

Given the upper bound on $d(\bar{\theta}, \theta^*)$, the location result then follows from Proposition 29 provided that $\mathrm{Err}_{h0}(\epsilon) \leq a$ for a constant $a \in [0, 1/2)$. The variance matrix result follows if $\mathrm{Err}_{h0}(\epsilon) \leq a/(1 + S_{1,a})$. ■

7.2 Proof of Theorem 15

We state and prove the following result which implies Theorem 15. See Appendix C.4 for details about how Proposition 21 implies Theorem 15. For $\delta \in (0, 1/7)$, define

$$\begin{aligned} \lambda_{11} &= \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{n}} + \frac{2\log(5p) + \log(\delta^{-1})}{n}, \\ \lambda_{12} &= 2C_{\mathrm{rad4}} \sqrt{\frac{\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}, \end{aligned}$$

where $C_{\mathrm{rad4}} = C_{\mathrm{sg6}} C_{\mathrm{rad3}}$, depending on universal constants C_{sg6} and C_{rad3} in Lemmas 70 and Corollary 82 in Appendix E. Denote

$$\mathrm{Err}_{h1}(n, p, \delta, \epsilon) = 3\epsilon + 2\sqrt{\epsilon/(n\delta)} + \lambda_{12} + \lambda_1,$$

where λ_1 is allowed to depend on λ_{11} in the following result.

Proposition 21 *Assume that* $\|\Sigma_*\|_{\max} \le M_1$ *and* $\epsilon \le 1/5$. *Let* $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ *be a solution to (20) with* $\lambda_1 \ge C_{\mathrm{sp13}} M_{11} \lambda_{11}$ *where* $M_{11} = M_1^{1/2}(M_1^{1/2} + 2\sqrt{2\pi})$ *and* $C_{\mathrm{sp13}} = (5/3)(C_{\mathrm{sp11}} \vee C_{\mathrm{sp12}})$, *depending on universal constants* C_{sp11} *and* C_{sp12} *in Lemma 30 in Appendix C. If* $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \le 1/5$ *and* $\mathrm{Err}_{h1}(n,p,\delta,\epsilon) \le a$ *for a constant* $a \in (0, 1/2)$, *then the following holds with probability at least* $1 - 7\delta$ *uniformly over contamination distribution* Q,

$$\|\hat{\mu} - \mu^*\|_\infty \le S_{4,a} \mathrm{Err}_{h1}(n,p,\delta,\epsilon),$$
$$\|\hat{\Sigma} - \Sigma^*\|_{\max} \le S_{8,a} \mathrm{Err}_{h1}(n,p,\delta,\epsilon),$$

where $S_{4,a} = (1 + \sqrt{2M_1 \log \frac{2}{1-2a}})/a$ *and* $S_{8,a} = 2M_1^{1/2} S_{6,a} + S_7(1 + S_{4,a} + S_{6,a})$ *with* $S_{6,a} = S_5(1 + S_{4,a}/2)$, $S_5 = 2\sqrt{2\pi}(1 - \mathrm{e}^{-2/M_1})^{-1}$, *and* $S_7 = 4\{(\frac{1}{\sqrt{2\pi M_1}} \mathrm{e}^{-1/(8M_1)}) \vee (1 - 2\mathrm{e}^{-1/(8M_1)})\}^{-2}$.

Remark 22 *In Proposition 21 as well as Proposition 23 for Theorem 11, the dependency of* $S_{4,a}$ *and* $S_{8,a}$ *on* M_1 *can be made explicit as follows. For fixed* $a \in (0, 1/2)$, *we have by direct calculation that* $\lim_{M_1 \to 0} S_{4,a} = 1/a$ *and* $\lim_{M_1 \to 0} S_{8,a} = 4 + 8\sqrt{2\pi} + (4\sqrt{2\pi} + 4)/a$. *Moreover,* $\lim_{M_1 \to \infty} S_{4,a}/{M_1}^{1/2} = \sqrt{2\log(2/(1-2a))}/a$ *and* $\lim_{M_1 \to \infty} S_{8,a}/M_1^{5/2} = 8\pi\sqrt{\log(2/(1-2a))}/a$, *that is,* $S_{4,a} = O(M_1^{1/2})$ *and* $S_{8,a} = O(M_1^{5/2})$ *as* $M_1 \to \infty$. *In addition,* λ_1 *in* $\mathrm{Err}_{h1}(n,p,\delta,\epsilon)$ *can be set to linearly depend on* M_1. *The overall dependency of our error rates on* M_1 *may potentially be improved. As mentioned in Section 5.1, our current analysis does not incorporate the scale transformation of real data, which may cause the sub-optimal dependency of* $S_{4,a}$ *and* $S_{8,a}$ *on* M_1.

Proof [Proof of Proposition 21] The main strategy of our proof is to show that the following inequalities hold with high probabilities,

$$d(\hat{\theta}, \theta^*) - \Delta_{12} \le \max_{\gamma \in \Gamma} \left\{ K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \mathrm{pen}_1(\gamma) \right\} \le \Delta_{11}, \tag{37}$$

where Δ_{11} and Δ_{12} are error terms, and $d(\theta^*, \hat{\theta})$ is a moment matching term, which under certain conditions delivers upper bounds, up to scaling constants, on the estimation errors to be controlled, $\|\hat{\mu} - \mu^*\|_\infty$ and $\|\hat{\Sigma} - \Sigma^*\|_{\max}$.

(Step 1) For upper bound in (37), we show that with probability at least $1 - 5\delta$,

$$\begin{aligned} &\max_{\gamma \in \Gamma} \left\{ K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \mathrm{pen}_1(\gamma) \right\} \\ &\le \max_{\gamma \in \Gamma} \left\{ K_{\mathrm{HG}}(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) - \lambda_1 \mathrm{pen}_1(\gamma) \right\} \qquad (38) \\ &\le \max_{\gamma \in \Gamma} \left\{ \Delta_{11} + \mathrm{pen}_1(\gamma)\tilde{\Delta}_{11} - \lambda_1 \mathrm{pen}_1(\gamma) \right\}. \qquad (39) \end{aligned}$$

Inequality (38) follows from the definition of $\hat{\theta}$. Inequality (39) follows from Proposition 52: it holds with probability at least $1 - 7\delta$ that for any $\gamma \in \Gamma$,

$$K_{\mathrm{HG}}(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \le \Delta_{11} + \mathrm{pen}_1(\gamma)\tilde{\Delta}_{11},$$

where $\Delta_{11} = 2(\epsilon + \sqrt{\epsilon/(n\delta)})$, $\tilde{\Delta}_{11} = C_{\text{sp13}} M_{11} \lambda_{11}$. From (38)–(39), the upper bound in (37) holds with probability at least $1 - 5\delta$, provided that the tuning parameter λ_1 is chosen such that $\lambda_1 \geq \tilde{\Delta}_{11}$.

(Step 2) For the lower bound in (37), we show that with probability at least $1 - 2\delta$,

$$\begin{aligned} &\max_{\gamma \in \Gamma} \left\{ K_{\text{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \, \text{pen}_1(\gamma) \right\} \\ \geq &\max_{\gamma \in \Gamma_0} \left\{ K_{\text{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \, \text{pen}_1(\gamma) \right\} \qquad (40) \\ \geq &\max_{\gamma \in \Gamma_0} \left\{ \text{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \text{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} - \tilde{\Delta}_{12} - \lambda_1. \qquad (41) \end{aligned}$$

Inequality (40) holds provided that Γ_0 is a subset of Γ.

Take $\Gamma_0 = \{\gamma \in \Gamma_{\text{rp}} : \gamma_0 = 0, \text{pen}_1(\gamma) = 1\}$, where Γ_{rp} is the subset of Γ associated with pairwise ramp functions as in the proof of Theorem 11. Inequality (41) follows from Proposition 53 because $h_{\gamma,\hat{\mu}}(x) \in [-1, 1]$ for $\gamma \in \Gamma_0$, and hence the hinge loss reduces to a moment matching term: it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_0$,

$$K_{\text{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \geq \left\{ \text{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \text{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} - \tilde{\Delta}_{12}$$

where $\tilde{\Delta}_{12} = \epsilon + \lambda_{12}$. From (40)–(41), the lower bound in (37) holds with probability at least $1 - 2\delta$, where $\Delta_{12} = \tilde{\Delta}_{12} + \lambda_1$ and $d(\hat{\theta}, \theta^*) = \max_{\gamma \in \Gamma_0} \{\text{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \text{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\}$.

(Step 3) We complete the proof by relating the moment matching term $d(\hat{\theta}, \theta^*)$ to the estimation error between $\hat{\theta}$ and θ^*. First, combining the lower and upper bounds in (37) shows that with probability at least $1 - 9\delta$,

$$\max_{\gamma \in \Gamma_{\text{rp}}, \text{pen}_1(\gamma) = 1} \left\{ \text{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \text{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} \leq \text{Err}_{h1}(n, p, \delta, \epsilon). \qquad (42)$$

where

$$\text{Err}_{h1}(n, p, \delta, \epsilon) = 3\epsilon + 2\sqrt{\epsilon/(n\delta)} + \lambda_{12} + \lambda_1.$$

The desired result then follows from Proposition 42: provided $\text{Err}_{h1}(n, p, \delta, \epsilon) \leq a$, inequality (42) implies that

$$\|\hat{\mu} - \mu^*\|_\infty \leq S_{4,a} \text{Err}_{h1}(n, p, \delta, \epsilon), \quad \|\hat{\Sigma} - \Sigma^*\|_{\max} \leq S_{8,a} \text{Err}_{h1}(n, p, \delta, \epsilon).$$

■

Acknowledgements

We thank the Action Editor and two referees for constructive comments. We also acknowledge the Office of Advanced Research Computing at Rutgers University for providing computing resources for the numerical studies reported here.

Appendix A. Additional information for simulation studies

A.1 Implementation of proposed methods

The detailed algorithm to implement our logit f-GANs and hinge GAN is shown as Algorithm 3. In our experiments, the default learning rates (α_d, α_g) are set to be $(0.002, 0.01)$ and increase fivefold when $p = 25$ or $n = 5000$, in the lower end of the range of p from 25 to 100 and n from 5000 to 50000 studied. The training steps (s_d, s_g) for the discriminator and the generator are $(20, 4)$, the mini-batch size is fixed to be 1000, and the total number

Algorithm 3: Penalized logit f-GAN or hinge GAN (in detail)

Require

1. A penalized GAN objective function $K(\theta, \gamma; \lambda)$ as in (25) for logit f-GAN or with K_f replaced by K_{HG} for hinge GAN.
2. Learning rates (α_d, α_g) for the discriminator and the generator;
3. Learning rate decay parameters (d, r) for the generator;
4. Numbers of training steps (s_d, s_g) for the discriminator and generator;
5. Base penalty level λ_0 so that $\lambda_1 = \lambda_0\sqrt{\log(p)/n}$, $\lambda_2 = \lambda_0\sqrt{p/n}$, and $\lambda_3 = \lambda_0\sqrt{p^2/n}$.
6. Mini-batch size m and number of epochs T.

Initialization

1. Initialize μ_0 by the median of X. Initialize discriminator intercept γ_0 by 0.01.
2. Initialize $\Sigma_0^{1/2}$ and the discriminator parameters (γ_1, γ_2) randomly by Xavier uniform (Glorot and Bengio, 2010).

for $t = 1 \dots T$ **do**
  for $u = 1 \dots T/m$ **do**
    Draw mini-batch $(x_1, \dots, x_m)$ from real data without replacement;
    for $s = 1 \dots s_d$ **do**
      Generate $(z_1, \dots, z_m)$ from $\mathrm{N}(0, I)$ and the fake data $\mu_{t-1} + \Sigma_{t-1}^{1/2} z_i$, $i = 1, \dots, m$;
      $g_\gamma \leftarrow \nabla_\gamma K(\theta_{t-1}, \gamma; \lambda)$; $g_\gamma \leftarrow g_\gamma / \|g_\gamma\|_2$;
      Update γ_t with gradient g_γ using the Adam algorithm (Kingma and Ba, 2015) with learning rate α_d.
    end
    for $s = 1 \dots s_g$ **do**
      Generate $(z_1, \dots, z_m)$ from $\mathrm{N}(0, I)$ and the fake data $\mu_{t-1} + \Sigma_{t-1}^{1/2} z_i$, $i = 1, \dots, m$;
      $g_\theta \leftarrow \nabla_\theta K(\theta, \gamma_t; \lambda)$; $g_\theta \leftarrow g_\theta / \|g_\theta\|_2$;
      Update θ_t with gradient g_θ using the Adam algorithm with learning rate α_g.
    end
    Decaying the generator learning rate: $\alpha_g \leftarrow r\alpha_g$ after every d epochs.
  end
end

of training epochs is set to be $T = 150 \times (50000/n)$ depending on n. We also decrease the learning rate of the generator as $\alpha_g \leftarrow r\alpha_g$ with $r = 0.5$ after each $10 \times (50000/n)$ epochs. This choice leads to stable convergence while keeping the running time relatively short.

For initialization of the variance matrix, we use a novel approach by treating the entries of $\Sigma_0^{1/2}$ as network weights and assigning uniform random numbers according to the Xavier uniform initialization in the neural network literature (Glorot and Bengio, 2010). This initialization scheme along with the Adam optimizer helps the generator accumulate momentum and overcome local minima issues. If initialized with Kendall's τ and MAD, it is possible that the generator may start near a generator local minimum and eventually become stuck there.

For implementation of rKL logit f-GAN, we modify the un-penalized objective

$$K_{\mathrm{rKL}}(P_*, P_\theta; h) = 1 - \mathrm{E}_{P_*} \mathrm{e}^{-h(x)} - \mathrm{E}_{P_\theta} h(x),$$

to

$$1 - \mathrm{E}_{P_*} \mathrm{e}^{-h(x)} + \max(-\mathrm{E}_{P_\theta} h(x), 9).$$

This modification caps the un-penalized rKL logit f-GAN objective by 10 and helps stabilize the initial steps of training where the fake data and real data, especially in the case of Cauchy contamination, can be separable. Despite the presence of an exponential term in the objective, the rKL logit f-GAN remains numerically stable because the trained discriminator h usually produces positive values on real data and the expectation of $\exp(-h(x))$ over real data is then upper bounded by 1. During the early training steps when the discriminator is relatively weak, any real data point x in the mini-batch that causes a much negative value $h(x)$ and an overflow of $\exp(-h(x))$ is dropped.

A.2 Tuning penalty levels

We conducted tuning experiments to identify base penalty levels λ_0 which are expected to work reasonably well in various settings for our logit f-GANs and hinge GAN, where the dependency on (p, n) is already absorbed in the penalty parameters $\lambda_1, \lambda_2, \lambda_3$. In the tuning experiments, we tried two contamination proportions ϵ and two choices of contamination distributions Q as described in Section 6.2. Results are collected from 20 repeated experiments on a grid of penalty levels for each method.

As shown in Figure 4, although the average estimation error varies as the contamination setting changes, there is a consistent and stable range of the penalty level λ_0 which leads to approximately the best performance for each method with L_1 penalty used. For L_2 penalized methods, although the pattern does not directly suggest a best choice of λ_0 in the range studied, the relative levels of estimation errors are less sensitive to the choice of λ_0. Hence we decide to use the same λ_0 for both L_1 and L_2 penalties. We manually pick $\lambda_0 = 0.1$ for the hinge GAN, $\lambda_0 = 0.025$ for the JS logit f-GAN, and $\lambda_0 = 0.3$ for the rKL logit f-GAN, which are then fixed in all subsequent simulations.

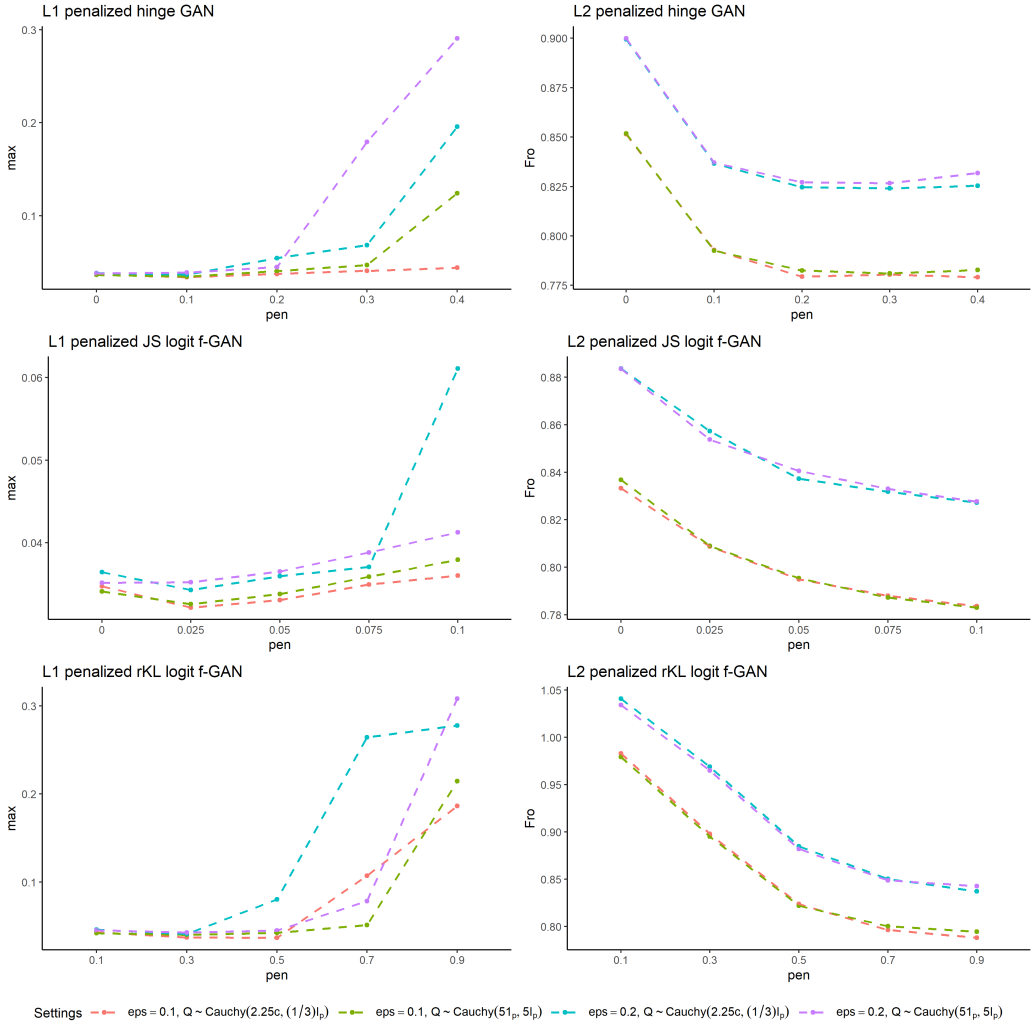

Figure 4: Average estimation errors against penalty levels. In this setting $p = 100$, $n = 10000$, $\epsilon \in \{0.1, 0.2\}$, and contamination distribution Q is either (A) Cauchy$(2.25c, (1/3)I_p)$ or (B) Cauchy$(51_p, 5I_p)$. Penalty levels for the hinge GAN are $\{0, 0.1, 0.2, 0.3, 0.4\}$, penalty levels for the JS logit f-GAN are $\{0, 0.025, 0.05, 0.075, 0.1\}$, and penalty levels for the rKL logit f-GAN are $\{0.1, 0.3, 0.5, 0.7, 0.9\}$.

n	hinge GAN	JS logit f-GAN	rKL logit f-GAN	GYZ JS-GAN	Kendall_MAD	Spearman_Qn	MCD	Tyler_M
				$Q \sim \text{Cauchy}\left(2.25c, \frac{1}{3}I_p\right)$				
5000	0.0762 (0.0066)	**0.0661 (0.0057)**	0.0713 (0.0054)	0.1123 (0.0229)	0.8251 (0.0225)	0.8801 (0.0145)	0.7016 (0.0207)	1.8392 (0.0616)
10000	0.0603 (0.0080)	**0.0491 (0.0059)**	0.0535 (0.0065)	0.3602 (0.5281)	0.7915 (0.0182)	0.8441 (0.0111)	0.6807 (0.0165)	1.8287 (0.0588)
20000	0.0394 (0.0047)	**0.0341 (0.0033)**	0.0343 (0.0034)	0.0527 (0.0096)	0.7514 (0.0122)	0.8297 (0.0055)	0.6510 (0.0065)	1.8045 (0.0368)
50000	0.0249 (0.0024)	**0.0217 (0.0031)**	0.0220 (0.0029)	0.0257 (0.0038)	0.7283 (0.0078)	0.8128 (0.0050)	0.6412 (0.0063)	1.7916 (0.0303)
				$Q \sim \text{Cauchy}(51_p, 5I_p)$				
5000	0.0845 (0.0061)	**0.0683 (0.0062)**	0.0857 (0.0068)	0.1053 (0.0172)	0.8826 (0.0264)	1.3862 (0.0241)	0.1114 (0.0094)	0.6212 (0.0214)
10000	0.0617 (0.0079)	**0.0507 (0.0058)**	0.0616 (0.0092)	0.0642 (0.0059)	0.8482 (0.0221)	1.3416 (0.0168)	0.0926 (0.0086)	0.6049 (0.0190)
20000	0.0433 (0.0045)	**0.0349 (0.0033)**	0.0385 (0.0043)	0.0483 (0.0113)	0.8071 (0.0127)	1.3132 (0.0104)	0.0821 (0.0040)	0.5850 (0.0118)
50000	0.0270 (0.0023)	**0.0216 (0.0024)**	0.0228 (0.0028)	0.0280 (0.0038)	0.7774 (0.0067)	1.2895 (0.0070)	0.0697 (0.0027)	0.5746 (0.0106)

Table 5: Comparison of existing methods and proposed L_1 penalized GAN methods ($p = 100$, $\epsilon = 0.2$, and varying n from 5000 to 50000). Estimation error of the variance matrix is reported in the maximum norm $\|\cdot\|_{\max}$.

n	hinge GAN	JS logit f-GAN	rKL logit f-GAN	GYZ JS-GAN	Kendall_MAD	Spearman_Qn	MCD	Tyler_M
				$Q \sim \text{Cauchy}\left(2.25c, \frac{1}{3}I_p\right)$				
5000	**1.6767 (0.0369)**	1.7287 (0.0363)	1.7815 (0.0412)	2.2490 (0.1139)	57.3156 (0.8202)	57.9363 (0.6875)	59.4896 (1.8053)	166.1099 (6.1982)
10000	**1.1616 (0.0194)**	1.1656 (0.0210)	1.1618 (0.0212)	27.4313 (50.7247)	57.0247 (0.6821)	58.1719 (0.5326)	58.8850 (1.2543)	166.5466 (5.6400)
20000	**0.8230 (0.0127)**	0.8257 (0.0108)	0.8324 (0.0139)	1.0726 (0.0509)	56.9405 (0.4234)	57.8273 (0.3415)	57.2124 (0.7063)	165.0762 (3.6783)
50000	**0.5164 (0.0092)**	0.5165 (0.0085)	0.5304 (0.0086)	0.6117 (0.0172)	56.8829 (0.2575)	57.9339 (0.1991)	57.2158 (0.5557)	165.1025 (3.0179)
				$Q \sim \text{Cauchy}(51_p, 5I_p)$				
5000	**1.6531 (0.0329)**	1.6603 (0.034)	1.6644 (0.0317)	2.2775 (0.1051)	53.3447 (1.0113)	65.3052 (1.2984)	1.8307 (0.0332)	55.3160 (2.1087)
10000	**1.1624 (0.0201)**	1.1686 (0.0203)	1.1767 (0.0213)	1.5363 (0.0371)	53.0505 (0.8515)	65.7633 (0.9984)	1.3728 (0.0393)	55.4997 (1.8886)
20000	**0.8252 (0.0114)**	0.8261 (0.0129)	0.8376 (0.0125)	1.0623 (0.0627)	52.9465 (0.5247)	65.1641 (0.5581)	1.0715 (0.0214)	54.9783 (1.2116)
50000	**0.5162 (0.0087)**	0.5171 (0.0089)	0.5330 (0.0088)	0.6380 (0.0651)	52.8773 (0.2932)	65.4056 (0.3857)	0.8515 (0.0126)	54.9977 (1.0236)

Table 6: Comparison of existing methods and proposed L_2 penalized GAN methods ($p = 100$, $\epsilon = 0.2$, and varying n from 5000 to 50000). Estimation error of the variance matrix is reported in the Frobenius norm $\|\cdot\|_{\mathrm{F}}$.

A.3 Error dependency on n and p

Tables 5–6 show the performance of various methods depending on sample size n for the two choices of contamination in Section 6.2. We fix the dimension $p = 100$ and $\epsilon = 0.2$ and increase n from 5000 to 50000. Tables 7–8 show how the performance of methods depending on sample size p for the two choices of contamination. We fix $\epsilon = 0.2$ and $n = 20000$ and increase p from 25 to 100. Estimation errors are measured in the maximum norm and the Frobenius norm.

For all methods considered, the estimation errors decrease as n increases except for JS-GAN in the first contamination setting (location $2.25c$). As can be seen in Tables 5-6, when $n = 10000$, we observe that 5 out of 20 runs appear to fail. In this setting, most outliers sit close to the uncontaminated data while a small number of outliers stretch to an extreme range. This makes it difficult for the discriminator to recognize both patterns and JS-GAN to perform satisfactorily, given that the discriminator objective surface is non-concave. The JS-GAN may require further tuning in this setting, but that is out of our scope.

It is also worth noting that with $\epsilon = 20\%$, the estimation errors of the coordinate-wise robust estimators (Kendall's τ and Spearman's ρ) show minimal decrease as n increases. This is because the error caused by the outliers tends to dominate the sampling variation, so that a 10-fold increase in n would not much reduce the overall error.

p	hinge GAN	JS logit f-GAN	rKL logit f-GAN	GYZ JS-GAN	Kendall_MAD	Spearman_Qn	MCD	Tyler_M
				$Q \sim \text{Cauchy}\left(2.25c, \frac{1}{3}I_p\right)$				
25	0.0452 (0.0052)	**0.0270 (0.0035)**	0.0271 (0.0034)	0.0406 (0.0092)	0.7429 (0.0126)	0.8184 (0.0127)	0.1632 (0.1263)	1.2278 (0.0210)
50	0.0355 (0.0055)	0.0309 (0.0046)	**0.0304 (0.0039)**	0.0410 (0.0092)	0.7505 (0.0097)	0.8216 (0.0076)	0.6566 (0.0113)	1.5793 (0.0327)
100	0.0394 (0.0047)	**0.0341 (0.0033)**	0.0343 (0.0034)	0.0527 (0.0096)	0.7514 (0.0122)	0.8297 (0.0055)	0.6510 (0.0065)	1.8045 (0.0368)
				$Q \sim \text{Cauchy}(51_p, 5I_p)$				
25	0.0531 (0.0057)	**0.0290 (0.0034)**	0.0364 (0.0044)	0.0413 (0.0125)	0.7968 (0.0130)	1.2938 (0.0136)	0.1348 (0.0076)	0.4320 (0.0080)
50	0.0423 (0.0066)	**0.0308 (0.0034)**	0.0350 (0.0043)	0.0399 (0.0051)	0.8040 (0.0092)	1.3053 (0.0177)	0.1021 (0.0068)	0.5190 (0.0120)
100	0.0433 (0.0045)	**0.0349 (0.0033)**	0.0385 (0.0043)	0.0483 (0.0113)	0.8071 (0.0127)	1.3132 (0.0104)	0.0821 (0.0040)	0.5850 (0.0118)

Table 7: Comparison of existing methods and proposed L_1 penalized GAN methods ($n = 20000$, $\epsilon = 0.2$, and varying p from 25 to 100). Estimation error of the variance matrix is reported in the maximum norm $\|\cdot\|_{\max}$.

p	hinge GAN	JS logit f-GAN	rKL logit f-GAN	GYZ JS-GAN	Kendall_MAD	Spearman_Qn	MCD	Tyler_M
				$Q \sim \text{Cauchy}\left(2.25c, \frac{1}{3}I_p\right)$				
25	0.2257 (0.0123)	**0.2015 (0.0143)**	0.2078 (0.0132)	0.2708 (0.0340)	14.2615 (0.1157)	14.6782 (0.0907)	1.4760 (3.2881)	26.4004 (0.4716)
50	**0.4049 (0.0102)**	0.4069 (0.0106)	0.4189 (0.0116)	0.4969 (0.0314)	28.4992 (0.2994)	29.1559 (0.1745)	29.1963 (0.4268)	70.8530 (1.5583)
100	**0.8230 (0.0127)**	0.8257 (0.0108)	0.8324 (0.0139)	1.0726 (0.0509)	56.9405 (0.4234)	57.8273 (0.3415)	57.2124 (0.7063)	165.0762 (3.6783)
				$Q \sim \text{Cauchy}(51_p, 5I_p)$				
25	0.2629 (0.0269)	**0.2337 (0.0236)**	0.2379 (0.0187)	0.2668 (0.0283)	13.4870 (0.1570)	17.2077 (0.1721)	0.7394 (0.0259)	9.7713 (0.1783)
50	0.4195 (0.0123)	**0.4135 (0.0108)**	0.4335 (0.0146)	0.4954 (0.0224)	26.6736 (0.3348)	33.3405 (0.3435)	0.8040 (0.0282)	24.1314 (0.5441)
100	**0.8252 (0.0114)**	0.8261 (0.0129)	0.8376 (0.0125)	1.0623 (0.0627)	52.9465 (0.5247)	65.1641 (0.5581)	1.0715 (0.0214)	54.9783 (1.2116)

Table 8: Comparison of existing methods and proposed L_2 penalized GAN methods ($n = 20000$, $\epsilon = 0.2$, and varying p from 25 to 100). Estimation error of the variance matrix is reported in the Frobenius norm $\|\cdot\|_{\text{F}}$.

As p increases, the estimation errors seem to be affected to a lesser extent when measured in the maximum norm. This is expected because an error rate $\sqrt{\log(p)/n}$ (ϵ term aside) has been established for our three L_1 penalized methods as well as Kendall's τ and Spearman's ρ (Loh and Tan, 2018). When measured in the Frobenius norm, the estimation errors go up as p increases, which is also expected.

In summary, our methods demonstrate remarkable consistency in handling various combinations of (p, n) for different types of contaminations. In contrast, the MCD and the two coordinate-wise robust estimators produce significantly different results when the contamination pattern changes. Although JS-GAN (Gao et al., 2020) achieves outstanding results in some cases, there are other cases where its performance is noticeably worse and less stable than our GAN methods with easy-to-train spline discriminators.

A.4 Illustration with the second contamination

Figure 5 shows the 95% Gaussian ellipses estimated for two selected coordinates, similarly as in Figure 1 but with two samples of size 20000 from a 100-dimensional Huber's contaminated Gaussian distributions based on the second contamination Q in Section 6.2. Comparison of the methods studied is qualitatively similar to that found in Figure 1. For completeness, the untruncated version of Figure 1 or 5 is presented in Figure 6 or 7 respectively. In each figure, only a random subsample of size 400 is included; otherwise the axes need to be of an even wider range to show the entire sample.

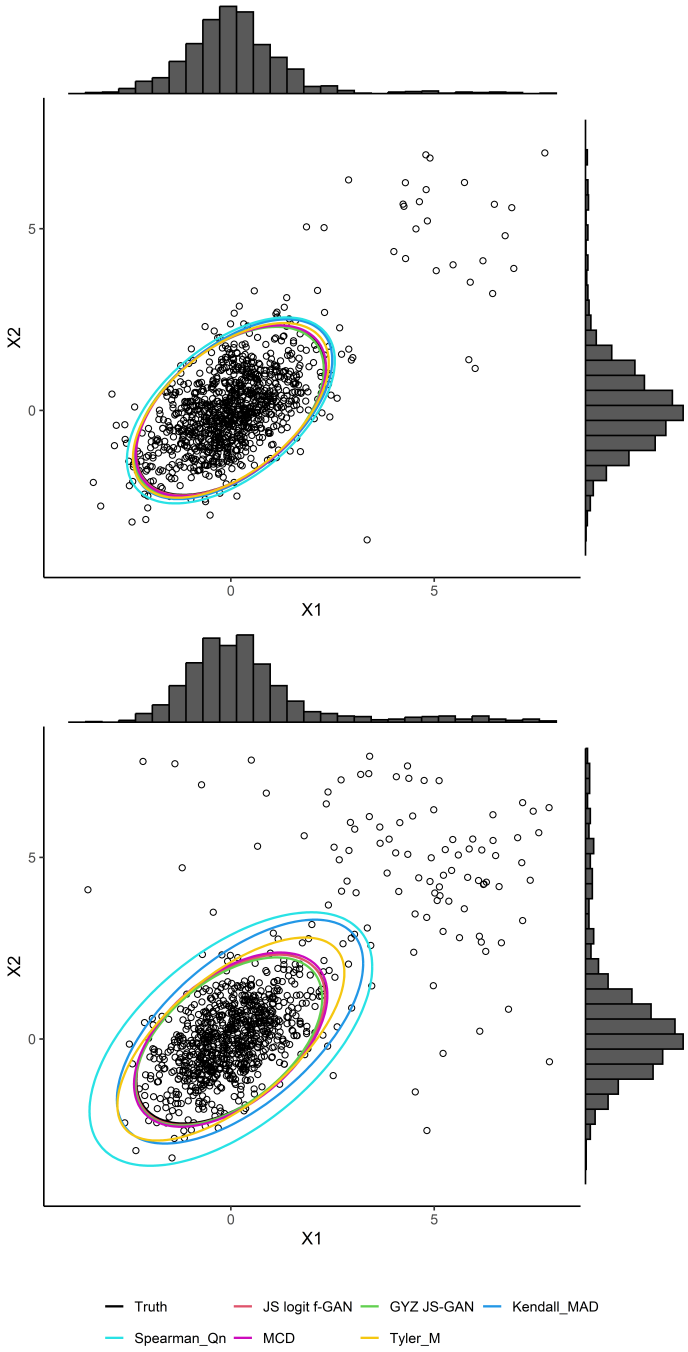

Figure 5: The estimated 95% Gaussian ellipses and observed marginal histograms for two selected coordinates, from contaminated data based on the second Cauchy contamination in Section 6.2 with $\epsilon = 5\%$ (top) or 20% (bottom). The data points are shown within the axis ranges $(-4, 8)$; see Figure 7 for untruncated plots.

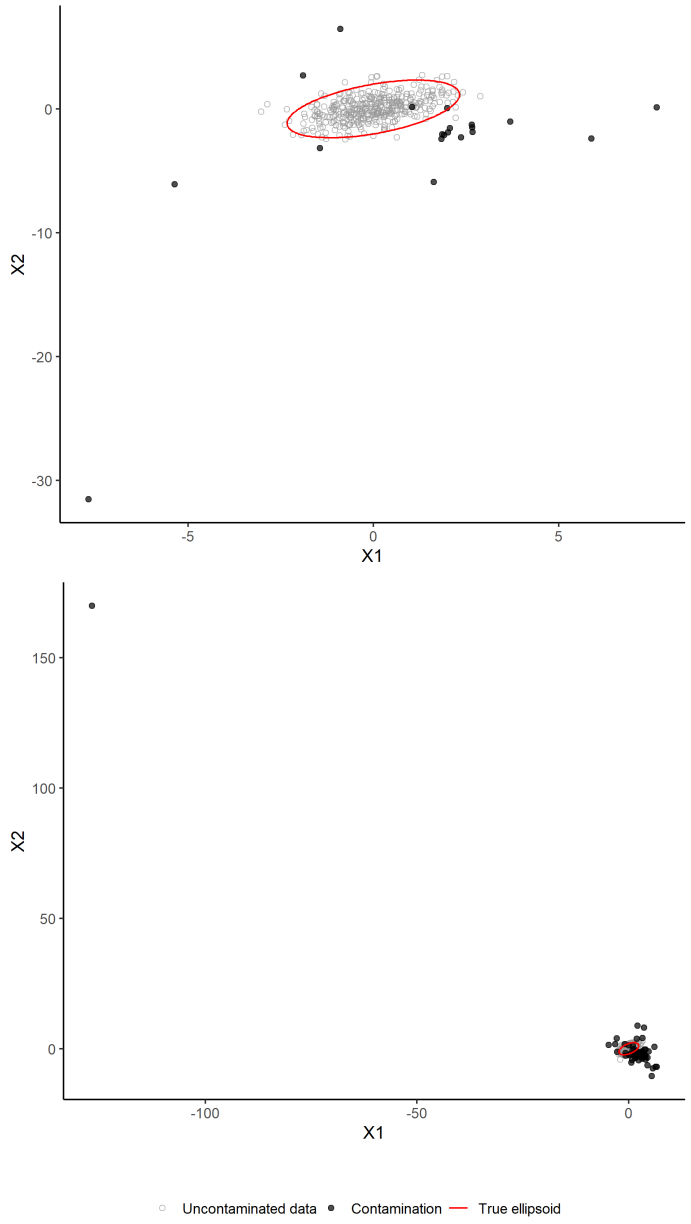

Figure 6: The untruncated version of Figure 1. Only the true 95% Gaussian ellipses are shown for two selected coordinates, from contaminated data based on the first Cauchy contamination in Section 6.2 with $\epsilon = 5\%$ (top) or 20% (bottom).

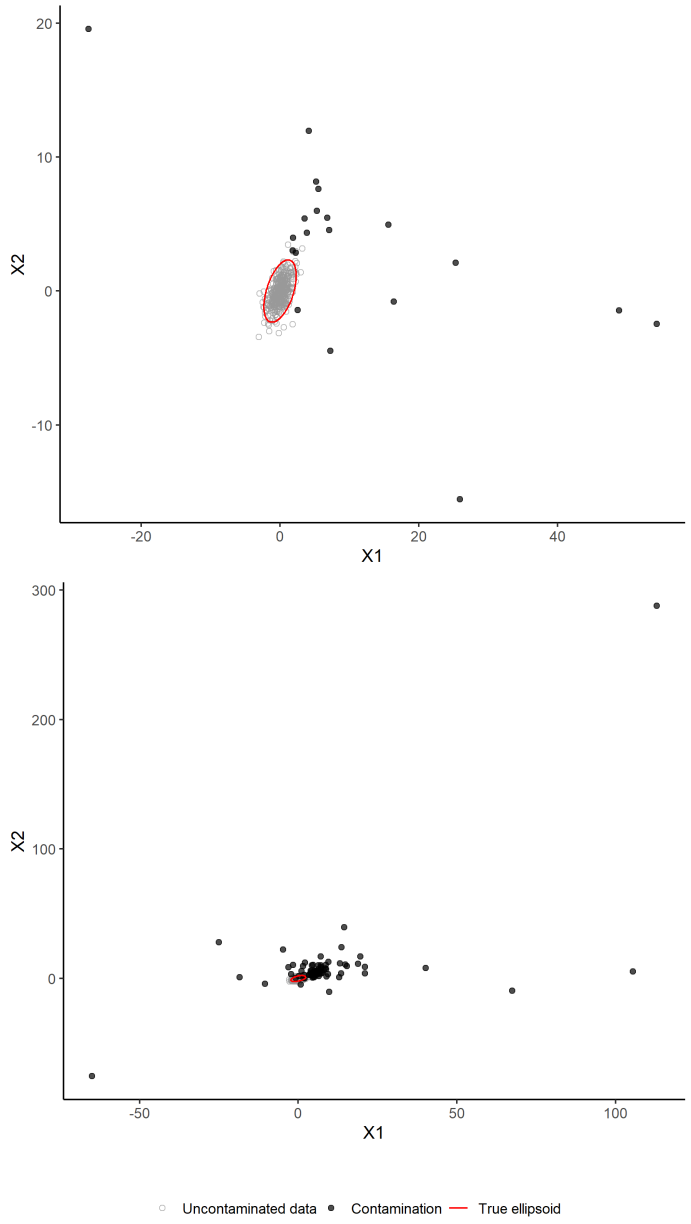

Figure 7: The untruncated version of Figure 5. Only the true 95% Gaussian ellipses are shown for two selected coordinates, from contaminated data based on the second Cauchy contamination in Section 6.2 with $\epsilon = 5\%$ (top) or 20% (bottom).

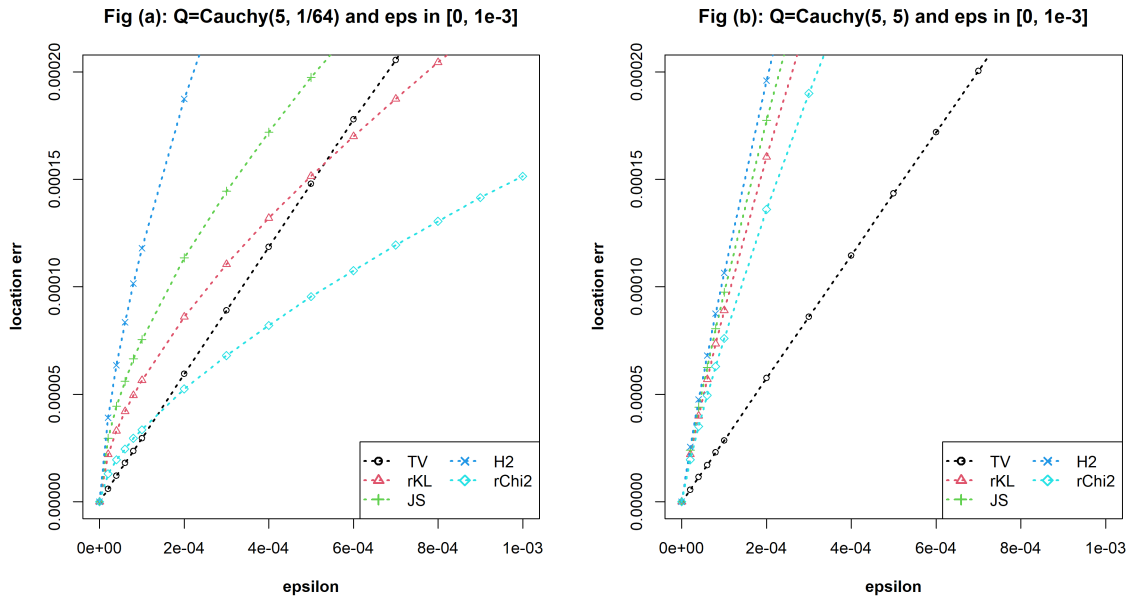

Figure 8: Comparison between two types of contamination in the error dependency on ϵ. Location error $|\bar{\mu} - \mu^*|$ against contamination fraction ϵ from 0 to 0.001, with P_{θ^*} being $\mathrm{N}(0,1)$. Figure (a): Non-overlapping contamination, Q being Cauchy$(5, 1/64)$; Figure (b): Overlapping contamination, Q being Cauchy$(5,5)$. The squared Hellinger and reverse χ^2 are denoted by H2 and rChi2 respectively.

A.5 Comparison of contamination settings

To provide further understanding of the worst-case contamination, we present in Figure 8 a comparison between two types of contamination for GANs at the population level, similarly to Figure 2. One type (non-overlapping contamination) may represent the worst-case contamination in terms of dependency on ϵ, where outliers do not overlap with the uncontaminated data. The errors from the robust f-divergence minimization exhibit square-root dependency on ϵ, whereas those from the TV minimization exhibit linear dependency on ϵ. The other type (overlapping contamination) is based on the second contamination used in our simulation studies. The errors from robust f-divergence and TV minimization appear to be linear in ϵ. Nevertheless, we also find that despite the worst-case dependency on ϵ, training of GANs with non-overlapping contaminations is numerically much easier than dealing with the two settings of overlapping contaminations in our simulation studies.

Appendix B. Main proofs of results

B.1 Proof of Theorem 11

We state and prove the following result which implies Theorem 11. For $b > 0$, define two factors $R_{2,b} = \sup_{|u| \le b} \frac{\mathrm{d}}{\mathrm{d}u} f'(\mathrm{e}^u)$ and $R_{3,b} = R_{31,b} + R_{32,b}$ with $R_{31,b} = \ sup_{|u| \le b} \frac{\mathrm{d}^2}{\mathrm{d}u^2}\{-f'(\mathrm{e}^u)\}$

and $R_{32,b} = \sup_{|u|\leq b} \frac{\mathrm{d}^2}{\mathrm{d}u^2} f^{\#}(\mathrm{e}^u)$. For $\delta \in (0,1)$, define

$$\lambda_{11} = \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{n}} + \frac{2\log(5p) + \log(\delta^{-1})}{n},$$
$$\lambda_{12} = C_{\mathrm{rad4}} \sqrt{\frac{4\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}},$$

where $C_{\mathrm{rad4}} = C_{\mathrm{sg6}} C_{\mathrm{rad3}}$, depending on universal constants C_{sg6} and C_{rad3} in Lemmas 70 and Corollary 82 in Appendix E. Denote

$$\mathrm{Err}_{f1}(n,p,\delta,\epsilon) = (f''(1))^{-1}\Big\{ - f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) \\ - f'(\mathrm{e}^{-b_1})\sqrt{\epsilon} + \frac{1}{2} R_{3,b_1}(\sqrt{\epsilon} + \sqrt{1/n}) + R_{2,b_1}\lambda_{12} + \lambda_1 \Big\},$$

where $b_1 = \sqrt{\epsilon} + \sqrt{1/n}$. Note that $R_{2,b}$, $R_{3,b}$ are bounded provided that b is bounded, because f is three-times continuously differentiable as required in Assumption 2.

Proposition 23 *Assume that $\|\Sigma^*\|_{\max} \leq M_1$, and f satisfies Assumptions 1–2. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (17) with $\lambda_1 \geq C_{\mathrm{sp13}} R_1 M_{11} \lambda_{11}$, where $M_{11} = M_1^{1/2}(M_1^{1/2} + 2\sqrt{2\pi})$ and $C_{\mathrm{sp13}} = (5/3)(C_{\mathrm{sp11}} \vee C_{\mathrm{sp12}})$, depending on universal constants C_{sp11} and C_{sp12} in Lemma 30 in Appendix C. If $\epsilon \leq 1/5$, $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \leq 1/5$, and $\mathrm{Err}_{f1}(n,p,\delta,\epsilon) \leq a$ for a constant $a \in (0,1/2)$, then we have that with probability at least $1 - 7\delta$,*

$$\|\hat{\mu} - \mu^*\|_\infty \leq S_{4,a} \mathrm{Err}_{f1}(n,p,\delta,\epsilon),$$
$$\|\hat{\Sigma} - \Sigma^*\|_{\max} \leq S_{8,a} \mathrm{Err}_{f1}(n,p,\delta,\epsilon),$$

where $S_{4,a} = (1 + \sqrt{2M_1 \log \frac{2}{1-2a}})/a$ and $S_{8,a} = 2M_1^{1/2} S_{6,a} + S_7(1 + S_{4,a} + S_{6,a})$ with $S_{6,a} = S_5(1 + S_{4,a}/2)$, $S_5 = 2\sqrt{2\pi}(1 - \mathrm{e}^{-2/M_1})^{-1}$, and $S_7 = 4\{(\frac{1}{\sqrt{2\pi M_1}} \mathrm{e}^{-1/(8M_1)}) \vee (1 - 2\mathrm{e}^{-1/(8M_1)})\}^{-2}$.

Proof [Proof of Proposition 23]

The main strategy of our proof is to show that the following inequalities hold with high probabilities,

$$d(\hat{\theta}, \theta^*) - \Delta_{12} \leq \max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \mathrm{pen}_1(\gamma) \right\} \leq \Delta_{11}, \tag{43}$$

where Δ_{11} and Δ_{12} are error terms, and $d(\theta^*, \hat{\theta})$ is a moment matching term, which under certain conditions delivers upper bounds, up to scaling constants, on the estimation errors to be controlled, $\|\hat{\mu} - \mu^*\|_\infty$ and $\|\hat{\Sigma} - \Sigma^*\|_{\max}$.

(Step 1) For the upper bound in (43), we show that with probability at least $1 - 5\delta$,

$$\max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \mathrm{pen}_1(\gamma) \right\}$$
$$\leq \max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) - \lambda_1 \mathrm{pen}_1(\gamma) \right\} \tag{44}$$
$$\leq \max_{\gamma \in \Gamma} \left\{ \Delta_{11} + \mathrm{pen}_1(\gamma)\tilde{\Delta}_{11} - \lambda_1 \mathrm{pen}_1(\gamma) \right\}. \tag{45}$$

Inequality (44) follows from the definition of $\hat{\theta}$. Inequality (45) follows from Proposition 33: it holds with probability at least $1-5\delta$ that for any $\gamma \in \Gamma$,

$$K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \le \Delta_{11} + \mathrm{pen}_1(\gamma)\tilde{\Delta}_{11},$$

where $\Delta_{11} = -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)})$, $\tilde{\Delta}_{11} = C_{\mathrm{sp13}} R_1 M_{11} \lambda_{11}$, and

$$\lambda_{11} = \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{n}} + \frac{2\log(5p) + \log(\delta^{-1})}{n}.$$

Note that λ_{11} is the same as in the proof of Theorem 15, and the above $\tilde{\Delta}_{11}$ differs from $\tilde{\Delta}_{11}$ in the proof of Theorem 15 only in the factor R_1. From (44)–(45), the upper bound in (43) holds with probability at least $1-5\delta$, provided that the tuning parameter λ_1 is chosen such that $\lambda_1 \ge \tilde{\Delta}_{11}$.

(Step 2) For the lower bound in (43), we show that with probability at least $1-2\delta$,

$$\begin{aligned}
&\max_{\gamma\in\Gamma} \left\{K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \,\mathrm{pen}_1(\gamma)\right\} \\
&\ge \max_{\gamma\in\Gamma_0} \left\{K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \,\mathrm{pen}_1(\gamma)\right\} \qquad (46)\\
&\ge \max_{\gamma\in\Gamma_0} f''(1)\left\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\right\} - \tilde{\Delta}_{12} - \lambda_1 b_1. \qquad (47)
\end{aligned}$$

Inequality (46) holds provided that Γ_0 is a subset of Γ. As a subset of the pairwise spline class $\mathcal{H}_{\mathrm{sp}}$, define a class of pairwise ramp functions, $\mathcal{H}_{\mathrm{rp}}$, such that each function in $\mathcal{H}_{\mathrm{rp}}$ can be expressed as, for $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$,

$$h_{\mathrm{rp},\beta,c}(x) = \beta_0 + \sum_{j=1}^{p} \beta_{1j}\,\mathrm{ramp}(x_j - c_j) + \sum_{1\le i\ne j\le p} \beta_{2,ij}\,\mathrm{ramp}(x_i)\mathrm{ramp}(x_j),$$

where $\mathrm{ramp}(t) = \frac{1}{2}(t+1)_+ - \frac{1}{2}(t-1)_+$ for $t \in \mathbb{R}$, $c = (c_1, \ldots, c_p)^{\mathrm{T}}$ with $c_j \in \{0,1\}$, and $\beta = (\beta_0, \beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}}$ with $\beta_1 = (\beta_{1j} : j = 1, \ldots, p)^{\mathrm{T}}$ and $\beta_2 = (\beta_{2,ij} : 1 \le i \ne j \le p)^{\mathrm{T}}$. For symmetry as in γ_2, assume that the coefficients in β_2 are symmetric, $\beta_{2,ij} = \beta_{2,ji}$ for any $i \ne j$. By the definition of $\mathrm{ramp}(\cdot)$, each function $h_{\mathrm{rp},\beta,c}(x)$ can be represented as $h_\gamma(x)$ in the spline class $\mathcal{H}_{\mathrm{sp}}$, where β and γ satisfy $\beta_0 = \gamma_0$, $\|\beta_1\|_1 = \|\gamma_1\|_1$, and $\|\beta_2\|_1 = \|\gamma_2\|_1$. Incidentally, this relationship also holds when symmetry is not imposed in the coefficients in γ_2 or in β_2. Denote as Γ_{rp} the subset of Γ such that $\mathcal{H}_{\mathrm{rp}} = \{h_\gamma(x) : \gamma \in \Gamma_{\mathrm{rp}}\}$.

Take $\Gamma_0 = \{\gamma \in \Gamma_{\mathrm{rp}} : \gamma_0 = 0, \mathrm{pen}_1(\gamma) = b_1\}$ for some fixed $b_1 > 0$. Inequality (47) follows from Proposition 37: it holds with probability at least $1-2\delta$ that for any $\gamma \in \Gamma_0$,

$$K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \ge f''(1)\left\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\right\} - \tilde{\Delta}_{12},$$

where $\tilde{\Delta}_{12} = -f'(\mathrm{e}^{-b_1})\epsilon + \frac{1}{2}b_1^2 R_{3,b_1} + b_1 R_{2,b_1}\lambda_{12}$, and

$$\lambda_{12} = C_{\mathrm{rad4}}\sqrt{\frac{4\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}.$$

Note that λ_{12} is the same as in the proof of Theorem 15. From (46)–(47), the lower bound in (43) holds with probability at least $1-2\delta$, where $\Delta_{12} = \tilde{\Delta}_{12} + \lambda_1 b_1$ and $d(\hat{\theta}, \theta^*) = f''(1) \max_{\gamma \in \Gamma_0} \{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \}$.

(Step 3) We complete the proof by choosing appropriate b_1 and relating the moment matching term $d(\hat{\theta}, \theta^*)$ to the estimation error between $\hat{\theta}$ and θ^*. First, due to the linearity of $h_{\gamma,\hat{\mu}}$ in γ, combining the lower and upper bounds in (43) shows that with probability at least $1-7\delta$,

$$\begin{aligned} & f''(1) b_1 \max_{\gamma \in \Gamma_{\mathrm{rp}}, \mathrm{pen}_1(\gamma)=1} \Big\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \Big\} \\ & \leq -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) - f'(\mathrm{e}^{-b_1})\epsilon + \frac{1}{2} b_1^2 R_{3,b_1} + b_1 R_{2,b_1} \lambda_{12} + \lambda_1 b_1. \end{aligned}$$

Taking $b_1 = \sqrt{\epsilon} + 1/\sqrt{n}$ in the preceding display and rearranging yields

$$\max_{\gamma \in \Gamma_{\mathrm{rp}}, \mathrm{pen}_1(\gamma)=1} \Big\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \Big\} \leq \mathrm{Err}_{f1}(n, p, \delta, \epsilon), \tag{48}$$

where

$$\begin{aligned} \mathrm{Err}_{f1}(n, p, \delta, \epsilon) = (f''(1))^{-1} \Big\{ & - f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) \\ & - f'(\mathrm{e}^{-b_1})\sqrt{\epsilon} + \frac{1}{2} R_{3,b_1}(\sqrt{\epsilon} + 1/\sqrt{n}) + R_{2,b_1}\lambda_{12} + \lambda_1 \Big\}. \end{aligned}$$

The desired result then follows from Proposition 42: provided $\mathrm{Err}_{f1}(n, p, \delta, \epsilon) \leq a$, inequality (48) implies that

$$\|\hat{\mu} - \mu^*\|_\infty \leq S_{4,a} \mathrm{Err}_{f1}(n, p, \delta, \epsilon), \quad \|\hat{\Sigma} - \Sigma^*\|_{\max} \leq S_{8,a} \mathrm{Err}_{f1}(n, p, \delta, \epsilon).$$

■

B.2 Proof of Theorem 12

We state and prove the following result which implies Theorem 12. For $b > 0$, define $R_{4,b} = \inf_{|u| \leq b} \frac{\mathrm{d}}{\mathrm{d}u} f^{\#}(\mathrm{e}^u)$, in addition to $R_{2,b}$ and $R_{3,b}$ as in Proposition 23. For $\delta \in (0,1)$, define

$$\begin{aligned} \lambda_{21} &= \sqrt{\frac{5p + \log(\delta^{-1})}{n}}, \quad \lambda_{22} = C_{\mathrm{rad5}} \sqrt{\frac{16p}{n}} + \sqrt{\frac{2p \log(\delta^{-1})}{n}}, \\ \lambda_{31} &= \lambda_{21} + \frac{5p + \log(\delta^{-1})}{n}, \quad \lambda_{32} = C_{\mathrm{rad5}} \sqrt{\frac{6(p-1)}{n}} + \sqrt{\frac{(p-1)\log(\delta^{-1})}{n}}. \end{aligned}$$

where $C_{\mathrm{rad5}} = C_{\mathrm{sg},12} C_{\mathrm{rad3}}$, depending on universal constants $C_{\mathrm{sg},12}$ and C_{rad3} in Lemmas 67 and Corollary 82 in Appendix E. Denote

$$\begin{aligned} \mathrm{Err}_{f2}(n, p, \delta, \epsilon) = (\sqrt{2} R_{4,b_2^\dagger})^{-1} \Big\{ & - f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) - f'(\mathrm{e}^{-b_2^\dagger})\sqrt{\epsilon} \\ & + 4 C_{\mathrm{sg},12}^2 M_2 R_{3,b_2^\dagger}(\sqrt{\epsilon} + \sqrt{1/(np)}) + R_{2,b_2^\dagger}\lambda_{22} + \lambda_2 \Big\}, \\ \mathrm{Err}_{f3}(n, p, \delta, \epsilon) = (2 R_{4,2b_3^\dagger})^{-1} \Big\{ & - f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) - f'(\mathrm{e}^{-2b_3^\dagger})\sqrt{\epsilon} \\ & + (80 C_{\mathrm{sg},12}^2 M_2) R_{3,2b_3^\dagger}(\sqrt{\epsilon} + \sqrt{1/(np)}) + R_{2,b_3^\dagger}\lambda_{32} + \lambda_3/\sqrt{p} \Big\}, \end{aligned}$$

where $b_2 = \sqrt{\epsilon} + \sqrt{1/(np)}$, $b_2^\dagger = b_2\sqrt{2p}$, $b_3 = \sqrt{\epsilon/p} + \sqrt{1/(np^2)}$, and $b_3^\dagger = b_2\sqrt{p(p-1)}$. Note that by the strict convexity and monotonicity of f as required in Assumption 1–2, we have that

$$R_{4,b} = \inf_{|u|\le b} \frac{\mathrm{d}}{\mathrm{d}u} f^{\#}(\mathrm{e}^u) = \inf_{|u|\le b} \frac{\mathrm{d}}{\mathrm{d}u} \left\{-f^*(f'(\mathrm{e}^u))\right\}$$

is bounded away from zero provided that b is bounded.

Proposition 24 *Assume that $\|\Sigma^*\|_{\mathrm{op}} \le M_2$, and f satisfies Assumptions 1–2. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (19) with*

$$\lambda_2 \ge (5/3) C_{\mathrm{sp}21} M_2^{1/2} R_1 \lambda_{21}, \quad \lambda_3/\sqrt{p} \ge (25\sqrt{5}/3) C_{\mathrm{sp}22} M_{21} R_1 \lambda_{31},$$

where $M_{21} = M_2^{1/2}(M_2^{1/2} + 2\sqrt{2\pi})$, $C_{\mathrm{sp}21} = \sqrt{2} C_{\mathrm{sg}7} C_{\mathrm{sg}5}$, and $C_{\mathrm{sp}22} = \sqrt{2/\pi} C_{\mathrm{sp}21} + C_{\mathrm{sg}8}$, depending on universal constants $C_{\mathrm{sg}5}$, $C_{\mathrm{sg}7}$, and $C_{\mathrm{sg}8}$ in Lemmas 69, 71, and 72 in Appendix E. If $\epsilon \le 1/5$, $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \le 1/5$, and $\mathrm{Err}_{f2}(n,p,\delta,\epsilon) \le a$ for a constant $a \in (0,1/2)$, then we have that with probability at least $1 - 8\delta$,

$$\|\hat{\mu} - \mu^*\|_2 \le S_{4,a} \mathrm{Err}_{f2}(n,p,\delta,\epsilon),$$
$$p^{-1/2}\|\hat{\Sigma} - \Sigma^*\|_{\mathrm{F}} \le S_{9,a} \mathrm{Err}_{f2}(n,p,\delta,\epsilon) + S_7 \mathrm{Err}_{f3}(n,p,\delta,\epsilon),$$

where $S_{9,a} = 2M_2^{1/2} S_{6,a} + \sqrt{2} S_7 (S_{4,a} + S_{6,a})$ and $(S_{4,a}, S_{6,a}, S_7)$ are defined as in Proposition 23 except with M_1 replaced by M_2 throughout.

Remark 25 *In Proposition 24 as well as Proposition 26 for Theorem 16, the dependency of $S_{4,a}$, S_7, and $S_{9,a}$ on M_2 can be made explicit as follows. For fixed $a \in (0,1/2)$, we have by direct calculation that $\lim_{M_2 \to 0} S_{4,a} = 1/a$, $\lim_{M_2 \to 0} S_7 = 4$, and $\lim_{M_2 \to 0} S_{9,a} = 16\sqrt{\pi} + (8\sqrt{\pi} + 4\sqrt{2})/a$. Moreover, $\lim_{M_2 \to \infty} S_{4,a}/{M_2}^{1/2} = \sqrt{2\log(2/(1-2a))}/a$, $\lim_{M_2 \to \infty} S_7/M_2 = 8\pi$ and $\lim_{M_2 \to \infty} S_{9,a}/M_1^{5/2} = 8\pi\sqrt{\log(2/(1-2a))}/a$, that is, $S_{4,a} = O(M_2^{1/2})$, $S_7 = O(M_2)$, and $S_{9,a} = O(M_2^{5/2})$ as $M_2 \to \infty$. In addition, λ_2 in $\mathrm{Err}_{f2}(n,p,\delta,\epsilon)$ can be set to linearly depend on $M_2^{1/2}$, and λ_3 in $\mathrm{Err}_{f3}(n,p,\delta,\epsilon)$ can be set to linearly depend on M_2. The overall dependency of our error rates on M_2 may potentially be improved, for a similar reason as discussed in Remark 22.*

Proof [Proof of Proposition 24] The main strategy of our proof is to show that the following inequalities hold with high probabilities,

$$d(\hat{\theta}, \theta^*) - \Delta_{22} \le \max_{\gamma \in \Gamma} \left\{K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_2 \,\mathrm{pen}_2(\gamma_1) - \lambda_3 \,\mathrm{pen}_2(\gamma_2)\right\} \le \Delta_{21}, \tag{49}$$

where Δ_{21} and Δ_{22} are error terms, and $d(\hat{\theta}, \theta^*)$ is a moment matching term, similarly as in the proof of Theorem 11. However, additional considerations are involved.

We split the proof into several steps. In Step 1, we derive the upper bound in (49) by exploiting two tuning parameters λ_2 and λ_3 associated with γ_1 and γ_2 respectively. In Steps 2 and 3, we derive the first version of the lower bound in (49) and then deduce upper bounds on $\|\hat{\mu} - \mu^*\|_2$ and $\|\hat{\sigma} - \sigma^*\|_2$, where $\hat{\sigma}$ or σ^* is the vector of standard deviations from $\hat{\Sigma}$ or Σ^* respectively. In Steps 4 and 5, we derive the second version of the lower bound in (49) and then deduce an upper bound on $\|\hat{\Sigma} - \Sigma^*\|_{\mathrm{F}}$.

(Step 1) For the upper bound in (49), we show that with probability at least $1-4\delta$,

$$\begin{aligned}
&\max_{\gamma\in\Gamma}\ \left\{K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) - \lambda_2\ \mathrm{pen}_2(\gamma_1) - \lambda_3\ \mathrm{pen}_2(\gamma_2)\right\}\\
\le &\max_{\gamma\in\Gamma}\ \{K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) - \lambda_2\ \mathrm{pen}_2(\gamma_1) - \lambda_3\ \mathrm{pen}_2(\gamma_2)\} \qquad (50)\\
\le &\max_{\gamma\in\Gamma}\ \left\{\Delta_{21} + \mathrm{pen}_2(\gamma_1)\tilde\Delta_{21} + \mathrm{pen}_2(\gamma_2)\tilde\Delta_{31} - \lambda_2\ \mathrm{pen}_2(\gamma_1) - \lambda_3\ \mathrm{pen}_2(\gamma_2)\right\}. \qquad (51)
\end{aligned}$$

Inequality (50) follows from the definition of $\hat\theta$. Inequality (51) follows from Proposition 44: it holds with probability at least $1-4\delta$ that for any $\gamma\in\Gamma_1$,

$$K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \le \Delta_{21} + \mathrm{pen}_2(\gamma_1)\tilde\Delta_{21} + \mathrm{pen}_2(\gamma_2)\sqrt{p}\tilde\Delta_{31},$$

where

$$\begin{aligned}
&\Delta_{21} = -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}), \quad \tilde\Delta_{21} = (5/3)C_{\mathrm{sp}21}M_2^{1/2}R_1\lambda_{21},\\
&\tilde\Delta_{31} = (25\sqrt{5}/3)C_{\mathrm{sp}22}M_{21}R_1\lambda_{31},
\end{aligned}$$

and

$$\lambda_{21} = \sqrt{\frac{5p+\log(\delta^{-1})}{n}}, \quad \lambda_{31} = \lambda_{21} + \frac{5p+\log(\delta^{-1})}{n}.$$

From (50)–(51), the upper bound in (49) holds with probability at least $1-4\delta$, provided that the tuning parameters λ_2 and λ_3 are chosen such that $\lambda_2 \ge \tilde\Delta_{21}$ and $\lambda_3 \ge \sqrt{p}\tilde\Delta_{31}$.

(Step 2) For the first version of the lower bound in (49), we show that with probability at least $1-2\delta$,

$$\begin{aligned}
&\max_{\gamma\in\Gamma}\ \left\{K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) - \lambda_2\ \mathrm{pen}_2(\gamma_1) - \lambda_3\ \mathrm{pen}_2(\gamma_2)\right\}\\
\ge &\max_{\gamma\in\Gamma_1}\ \left\{K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) - \lambda_2\ \mathrm{pen}_2(\gamma_1)\right\} \qquad (52)\\
\ge &\max_{\gamma\in\Gamma_{10}}\ \left\{K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) - \lambda_2\ \mathrm{pen}_2(\gamma_1)\right\} \qquad (53)\\
\ge &\max_{\gamma\in\Gamma_{10}}\ R_{4,b_2^\dagger}\left\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x) - \mathrm{E}_{P_{\hat\theta}} h_{\gamma,\hat\mu}(x)\right\} - \tilde\Delta_{22} - \lambda_2 b_2, \qquad (54)
\end{aligned}$$

where $\Gamma_1 = \{(\gamma_0, \gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}})^{\mathrm{T}} : \gamma_2 = 0\}$. Inequality (52) follows because Γ_1 is a subset of Γ such that $\gamma_2 = 0$ and hence $\mathrm{pen}_2(\gamma_2) = 0$ for $\gamma\in\Gamma_1$. Inequality (53) holds provided that Γ_{10} is a subset of Γ_1. As a subset of the main-effect spline class $\mathcal{H}_{\mathrm{sp}1}$, define a main-effect ramp class, $\mathcal{H}_{\mathrm{rp}1}$, such that each function in $\mathcal{H}_{\mathrm{rp}1}$ can be expressed as, for $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$,

$$h_{\mathrm{rp}1,\beta,c}(x) = \beta_0 + \sum_{j=1}^{p} \beta_{1j}\mathrm{ramp}(x_j - c_j),$$

where $\mathrm{ramp}(t) = \frac{1}{2}(t+1)_+ - \frac{1}{2}(t-1)_+$ for $t\in\mathbb{R}$, $c = (c_1, \ldots, c_p)^{\mathrm{T}}$ with $c_j \in \{0,1\}$, and $\beta = (\beta_0, \beta_1^{\mathrm{T}})^{\mathrm{T}}$ with $\beta_1 = (\beta_{11}, \ldots, \beta_{1p})^{\mathrm{T}}$. Only the main-effect ramp functions are included, while the interaction ramp functions are excluded, in $h_{\mathrm{rp}1,\beta,c}(x)$. By the definition of $\mathrm{ramp}(\cdot)$, each function $h_{\mathrm{rp}1,\beta,c}(x)$ can be represented as $h_\gamma(x) \in \mathcal{H}_{\mathrm{sp}1}$ with $\gamma = (\gamma_0, \gamma_1^{\mathrm{T}})^{\mathrm{T}} \in \Gamma_{\mathrm{rp}1}$,

such that β and γ satisfy $\beta_0 = \gamma_0$ and $\|\beta_1\|_2 = \sqrt{2}\|\gamma_1\|_2$. For example, for ramp$(x_1)$, the associated norms are $\|\beta_1\|_2 = 1$ and $\|\gamma_1\|_2 = \sqrt{1/2}$. Denote as Γ_{rp1} the subset of Γ_1 such that $\mathcal{H}_{\mathrm{rp1}} = \{h_\gamma(x) : \gamma \in \Gamma_{\mathrm{rp1}}\}$.

Take $\Gamma_{10} = \{\gamma \in \Gamma_{\mathrm{rp1}} : \mathrm{pen}_2(\gamma) = b_2, \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0, \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \leq 0\}$ for some fixed $b_2 > 0$. Inequality (54) follows from Proposition 47: it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{10}$,

$$K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \geq R_{4,b_2^\dagger} \Big\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \Big\} - \tilde{\Delta}_{22},$$

where $b_2^\dagger = b_2\sqrt{2p}$, $\tilde{\Delta}_{22} = -f'(\mathrm{e}^{-b_2^\dagger})\epsilon + 4C_{\mathrm{sg},12}^2 M_2 b_2^2 R_{3,b_2^\dagger} + b_2 R_{2,b_2^\dagger}\lambda_{22}$, and

$$\lambda_{22} = C_{\mathrm{rad}5}\sqrt{\frac{16p}{n}} + \sqrt{\frac{2p\log(\delta^{-1})}{n}}.$$

From (53)–(54), the lower bound in (49) holds with probability at least $1 - 2\delta$, where $\Delta_{22} = \tilde{\Delta}_{22} + \lambda_2 b_2$ and $d(\hat{\theta}, \theta^*) = R_{4,b_2^\dagger}\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\}$.

(Step 3) We deduce upper bounds on $\|\hat{\mu} - \mu^*\|_2$ and $\|\hat{\sigma} - \sigma^*\|_2$, by choosing appropriate b_2 and relating the moment matching term $d(\hat{\theta}, \theta^*)$ to the estimation errors. First, combining the upper bound in (49) from Step 1 and the lower bound from Step 2 shows that with probability at least $1 - 6\delta$,

$$\begin{aligned} &R_{4,b_2^\dagger} b_2 \max_{\gamma \in \Gamma_{\mathrm{rp1}}, \mathrm{pen}_2(\gamma) = 1} \Big\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \Big\} \\ &\leq -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) - f'(\mathrm{e}^{-b_2^\dagger})\epsilon + 4C_{\mathrm{sg},12}^2 M_2 b_2^2 R_{3,b_2^\dagger} + b_2 R_{2,b_2^\dagger}\lambda_{22} + \lambda_2 b_2. \end{aligned}$$

Taking $b_2 = \sqrt{\epsilon} + \sqrt{1/(np)}$ in the preceding display and rearranging yields

$$\max_{\gamma \in \Gamma_{\mathrm{rp1}}, \mathrm{pen}_2(\gamma) = \sqrt{1/2}} \Big\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \Big\} \leq \mathrm{Err}_{f2}(n, p, \delta, \epsilon), \tag{55}$$

where

$$\begin{aligned} \mathrm{Err}_{f2}(n, p, \delta, \epsilon) = (\sqrt{2} R_{4,b_2^\dagger})^{-1} \Big\{ &- f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) \\ &- f'(\mathrm{e}^{-b_2^\dagger})\sqrt{\epsilon} + 4C_{\mathrm{sg},12}^2 M_2 R_{3,b_2^\dagger}(\sqrt{\epsilon} + \sqrt{1/(np)}) + R_{2,b_2^\dagger}\lambda_{22} + \lambda_2 \Big\}. \end{aligned}$$

The error bounds for $(\hat{\mu}, \hat{\sigma})$ then follows from Proposition 48: provided $\mathrm{Err}_{f2}(n, p, \delta, \epsilon) \leq a$, inequality (55) implies that

$$\|\hat{\mu} - \mu^*\|_2 \leq S_{4,a}\mathrm{Err}_{f2}(n, p, \delta, \epsilon), \tag{56}$$

$$\|\hat{\sigma} - \sigma^*\|_2 \leq S_{6,a}\mathrm{Err}_{f2}(n, p, \delta, \epsilon). \tag{57}$$

(Step 4) For the second version of the lower bound in (49), we show that with probability at least $1-2\delta$,

$$\begin{aligned}
&\max_{\gamma\in\Gamma}\ \left\{K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) - \lambda_2\ \mathrm{pen}_2(\gamma_1) - \lambda_3\ \mathrm{pen}_2(\gamma_2)\right\}\\
&\geq \max_{\gamma\in\Gamma_2}\ \left\{K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) - \lambda_3\ \mathrm{pen}_2(\gamma_2)\right\} \qquad (58)\\
&\geq \max_{\gamma\in\Gamma_{20}}\ \left\{K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) - \lambda_3\ \mathrm{pen}_2(\gamma_2)\right\} \qquad (59)\\
&\geq \max_{\gamma\in\Gamma_{20}}\ f''(1)\left\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x) - \mathrm{E}_{P_{\hat\theta}} h_{\gamma,\hat\mu}(x)\right\} - \tilde\Delta_{32} - \lambda_3 b_3, \qquad (60)
\end{aligned}$$

where $\Gamma_2 = \{(\gamma_0, \gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}})^{\mathrm{T}} : \gamma_1 = 0\}$. Inequality (58) follows because Γ_2 is a subset of Γ such that $\gamma_1 = 0$ and hence $\mathrm{pen}_2(\gamma_1) = 0$ for $\gamma \in \Gamma_2$. Inequality (59) holds provided that Γ_{20} is a subset of Γ_2. As a subset of the interaction spline class $\mathcal{H}_{\mathrm{sp2}}$, define an interaction ramp class, $\mathcal{H}_{\mathrm{rp2}}$, such that each function in $\mathcal{H}_{\mathrm{rp2}}$ can be expressed as, for $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$,

$$h_{\mathrm{rp2},\beta}(x) = \beta_0 + \sum_{1\leq i\neq j\leq p} \beta_{2,ij}\mathrm{ramp}(x_i)\mathrm{ramp}(x_j),$$

where $\mathrm{ramp}(t) = \frac{1}{2}(t+1)_+ - \frac{1}{2}(t-1)_+$ for $t \in \mathbb{R}$, and $\beta = (\beta_0, \beta_2^{\mathrm{T}})^{\mathrm{T}}$ with $\beta_2 = (\beta_{2,ij} : 1 \leq i \neq j \leq p)^{\mathrm{T}}$. In contrast with the function $h_{\mathrm{rp1},\beta,c}(x)$ in $\mathcal{H}_{\mathrm{sp1}}$, only the interaction ramp functions are included, while the main-effect ramp functions are excluded, in $h_{\mathrm{rp2},\beta}(x)$. For symmetry as in γ_2, assume that the coefficients in β_2 are symmetric, $\beta_{2,ij} = \beta_{2,ji}$ for any $i \neq j$. By the definition of $\mathrm{ramp}(\cdot)$, each function $h_{\mathrm{rp2},\beta}(x)$ can be represented as $h_\gamma(x) \in \mathcal{H}_{\mathrm{sp2}}$ with $\gamma = (\gamma_0, \gamma_2^{\mathrm{T}})^{\mathrm{T}} \in \Gamma_{\mathrm{rp2}}$, such that β and γ satisfy $\beta_0 = \gamma_0$ and $\|\beta_2\|_2 = 2\|\gamma_2\|_2$. For example, for $\mathrm{ramp}(x_1)\mathrm{ramp}(x_2)$, the associated norms are $\|\beta_2\|_2 = 1$ and $\|\gamma_2\|_2 = 1/2$. Denote as Γ_{rp2} the subset of Γ_2 such that $\mathcal{H}_{\mathrm{rp2}} = \{h_\gamma(x) : \gamma \in \Gamma_{\mathrm{rp2}}\}$.

Take $\Gamma_{20} = \{\gamma \in \Gamma_{\mathrm{rp2}} : \mathrm{pen}_2(\gamma) = b_3, \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x) = 0, \mathrm{E}_{P_{\hat\theta}} h_{\gamma,\hat\mu}(x) \leq 0\}$ for some fixed $b_3 > 0$. Inequality (60) follows from Proposition 49: it holds with probability at least $1-2\delta$ that for any $\gamma \in \Gamma_{20}$,

$$K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) \geq R_{4,2b_3^\dagger}\left\{\mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat\theta}} h(x)\right\} - \tilde\Delta_{32},$$

where $b_3^\dagger = b_3\sqrt{p(p-1)}$, $\tilde\Delta_{32} = -f'(\mathrm{e}^{-2b_3^\dagger})\epsilon + (80C_{\mathrm{sg},12}^2 M_2)pb_3^2 R_{3,2b_3^\dagger} + \sqrt{p}b_3 R_{2,b_3^\dagger}\lambda_{32}$, and

$$\lambda_{32} = C_{\mathrm{rad}4}\sqrt{\frac{6(p-1)}{n}} + \sqrt{\frac{(p-1)\log(\delta^{-1})}{n}}.$$

From (59)–(60), the lower bound in (49) holds with probability at least $1-2\delta$, where $\Delta_{22} = \tilde\Delta_{32} + \lambda_3 b_3$ and $d(\hat\theta, \theta^*) = R_{4,2b_3^\dagger} \max_{\gamma\in\Gamma_{20}}\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x) - \mathrm{E}_{P_{\hat\theta}} h_{\gamma,\hat\mu}(x)\}$.

(Step 5) We deduce an upper bound on $\|\hat\Sigma - \Sigma^*\|_{\mathrm{F}}$, by choosing appropriate b_3 and relating the moment matching term $d(\hat\theta, \theta^*)$ to the estimation error. First, combining the upper bound in (49) from Step 1 and the lower bound from Step 4 shows that with probability $1-6\delta$,

$$\begin{aligned}
&R_{4,2b_3^\dagger} b_3 \max_{\gamma\in\Gamma_{\mathrm{rp2}},\mathrm{pen}_2(\gamma)=1}\left\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x) - \mathrm{E}_{P_{\hat\theta}} h_{\gamma,\hat\mu}(x)\right\}\\
&\leq -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) - f'(\mathrm{e}^{-2b_3^\dagger})\epsilon + (80C_{\mathrm{sg},12}^2 M_2)pb_3^2 R_{3,2b_3^\dagger} + \sqrt{p}b_3 R_{2,b_3^\dagger}\lambda_{32} + \lambda_3 b_3.
\end{aligned}$$

Taking $b_3 = \sqrt{\epsilon/p} + \sqrt{1/(np^2)}$ in the preceding display and rearranging yields

$$\max_{\gamma\in\Gamma_{\mathrm{rp2}},\mathrm{pen}_2(\gamma)=1/2} \Big\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \Big\} \leq \sqrt{p}\, \mathrm{Err}_{f3}(n,p,\delta,\epsilon), \tag{61}$$

where

$$\begin{aligned}\mathrm{Err}_{f3}(n,p,\delta,\epsilon) = (2R_{4,2b_3^\dagger})^{-1}\Big\{ &- f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) \\ &- f'(\mathrm{e}^{-2b_3^\dagger})\sqrt{\epsilon} + (80C_{\mathrm{sg},12}^2 M_2) R_{3,2b_3^\dagger}(\sqrt{\epsilon} + \sqrt{1/(np)}) + R_{2,b_3^\dagger}\lambda_{32} + \lambda_3/\sqrt{p}\Big\}.\end{aligned}$$

The error bound for $\hat{\Sigma}$ then follows from Proposition 50: inequality (61) together with the error bounds (56)–(57) implies that

$$\begin{aligned}\frac{1}{\sqrt{p}}\|\hat{\Sigma} - \Sigma^*\|_{\mathrm{F}} &\leq 2M_2^{1/2}\|\hat{\sigma} - \sigma^*\|_2 + S_7\Big\{\sqrt{2}\Delta_{\hat{\mu},\hat{\sigma}} + \mathrm{Err}_{f3}(n,p,\delta,\epsilon)\Big\} \\ &\leq S_{9,a}\mathrm{Err}_{f2}(n,p,\delta,\epsilon) + S_7\mathrm{Err}_{f3}(n,p,\delta,\epsilon),\end{aligned}$$

where $\Delta_{\hat{\mu},\hat{\sigma}} = (\|\hat{\mu} - \mu^*\|_2^2 + \|\hat{\sigma} - \sigma^*\|_2^2)^{1/2}$ and $S_{9,a} = 2M_2^{1/2}S_{6,a} + \sqrt{2}S_7(S_{4,a} + S_{6,a})$. ■

B.3 Proof of Theorem 16

We state and prove the following result which implies Theorem 16. For $\delta \in (0,1)$, define $(\lambda_{21}, \lambda_{31}, \lambda_{22}, \lambda_{32})$ the same as in Sections B.1 and B.2. Denote

$$\begin{aligned}\mathrm{Err}_{h2}(n,p,\delta,\epsilon) &= 3\epsilon(2p)^{1/2} + 2\sqrt{2p\epsilon/(n\delta)} + \lambda_2 + \lambda_{22}, \\ \mathrm{Err}_{h3}(n,p,\delta,\epsilon) &= 3\epsilon\sqrt{p-1} + 2\sqrt{\epsilon(p-1)/(n\delta)} + \lambda_{32}/2 + (25\sqrt{5}/6)C_{\mathrm{sp}22}M_{21}\lambda_{31}.\end{aligned}$$

Proposition 26 *Assume that $\|\Sigma^*\|_{\mathrm{op}} \leq M_2$. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (21) with $\lambda_3/\sqrt{p} \geq (25\sqrt{5}/3)C_{\mathrm{sp}22}M_{21}\lambda_{31}$ and $\lambda_2 \geq (5/3)C_{\mathrm{sp}21}M_2^{1/2}\lambda_{21}$, where M_{21}, $C_{\mathrm{sp}21}$, and $C_{\mathrm{sp}22}$ are defined as in Proposition 24. If $\epsilon \leq 1/5$, $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \leq 1/5$ and $\mathrm{Err}_{h2}(n,p,\delta,\epsilon) \leq a$ for a constant $a \in (0,1/2)$, then we have that with probability at least $1 - 8\delta$,*

$$\begin{aligned}\|\hat{\mu} - \mu^*\|_2 &\leq S_{4,a}\mathrm{Err}_{h2}(n,p,\delta,\epsilon), \\ p^{-1/2}\|\hat{\Sigma} - \Sigma^*\|_{\mathrm{F}} &\leq S_{9,a}\mathrm{Err}_{h2}(n,p,\delta,\epsilon) + S_7\mathrm{Err}_{h3}(n,p,\delta,\epsilon),\end{aligned}$$

where $(S_{4,a}, S_{6,a}, S_7, S_{9,a})$ are defined as in Proposition 24.

Proof [Proof of Proposition 26]

The main strategy of our proof is to show that the following inequalities hold with high probabilities,

$$d(\hat{\theta},\theta^*) - \Delta_{22} \leq \max_{\gamma\in\Gamma} \big\{K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_2\,\mathrm{pen}_2(\gamma_1) - \lambda_3\,\mathrm{pen}_2(\gamma_2)\big\} \leq \Delta_{21}, \tag{62}$$

where Δ_{21} and Δ_{22} are error terms, and $d(\hat{\theta},\theta^*)$ is a moment matching term, similarly as in the proof of Theorem 15. However, additional considerations are involved.

(Step 1) For the upper bound in (62), we show that with probability at least $1-4\delta$,

$$\begin{aligned}
&\max_{\gamma\in\Gamma}\ \left\{K_{\mathrm{HG}}(P_n,P_{\hat\theta};h_{\gamma,\hat\mu})-\lambda_2\ \mathrm{pen}_2(\gamma_1)-\lambda_3\ \mathrm{pen}_2(\gamma_2)\right\}\\
\le&\max_{\gamma\in\Gamma}\ \left\{K_{\mathrm{HG}}(P_n,P_{\theta^*};h_{\gamma,\mu^*})-\lambda_2\ \mathrm{pen}_2(\gamma_1)-\lambda_3\ \mathrm{pen}_2(\gamma_2)\right\} &(63)\\
\le&\max_{\gamma\in\Gamma}\ \left\{\Delta_{21}+\mathrm{pen}_2(\gamma_1)\tilde\Delta_{21}+\mathrm{pen}_2(\gamma_2)\tilde\Delta_{31}-\lambda_2\ \mathrm{pen}_2(\gamma_1)-\lambda_3\ \mathrm{pen}_2(\gamma_2)\right\}. &(64)
\end{aligned}$$

Inequality (63) follows from the definition of $\hat\theta$. Inequality (64) follows from Proposition 54: it holds with probability at least $1-4\delta$ that for any $\gamma\in\Gamma_1$,

$$K_{\mathrm{HG}}(P_n,P_{\theta^*};h_{\gamma,\mu^*})\le\Delta_{21}+\mathrm{pen}_2(\gamma_1)\tilde\Delta_{21}+\mathrm{pen}_2(\gamma_2)\sqrt{p}\tilde\Delta_{31},$$

where

$$\Delta_{21}=2(\epsilon+\sqrt{\epsilon/(n\delta)}),\quad \tilde\Delta_{21}=(5/3)C_{\mathrm{sp}21}M_2^{1/2}\lambda_{21},\quad \tilde\Delta_{31}=(25\sqrt{5}/3)C_{\mathrm{sp}22}M_{21}\lambda_{31},$$

and λ_{21} and λ_{31} are the same as in the proof of Theorem 12. Note that $\tilde\Delta_{21}$ and $\tilde\Delta_{31}$ differ from those in the proof of Theorem 12 only in that R_1 is removed. From (63)–(64), the upper bound in (62) holds with probability at least $1-4\delta$, provided that the tuning parameters λ_2 and λ_3 are chosen such that $\lambda_2\ge\tilde\Delta_{21}$ and $\lambda_3\ge\sqrt{p}\tilde\Delta_{31}$.

(Step 2) For the first version of the lower bound in (62), we show that with probability at least $1-2\delta$,

$$\begin{aligned}
&\max_{\gamma\in\Gamma}\ \left\{K_{\mathrm{HG}}(P_n,P_{\hat\theta};h_{\gamma,\hat\mu})-\lambda_2\ \mathrm{pen}_2(\gamma_1)-\lambda_3\ \mathrm{pen}_2(\gamma_2)\right\}\\
\ge&\max_{\gamma\in\Gamma_1}\ \left\{K_{\mathrm{HG}}(P_n,P_{\hat\theta};h_{\gamma,\hat\mu})-\lambda_2\ \mathrm{pen}_2(\gamma_1)\right\} &(65)\\
\ge&\max_{\gamma\in\Gamma_{10}}\ \left\{K_{\mathrm{HG}}(P_n,P_{\hat\theta};h_{\gamma,\hat\mu})\right\}-\lambda_2\ (2p)^{-1/2} &(66)\\
\ge&\max_{\gamma\in\Gamma_{10}}\ \left\{\mathrm{E}_{P_{\theta^*}}h_{\gamma,\hat\mu}(x)-\mathrm{E}_{P_{\hat\theta}}h_{\gamma,\hat\mu}(x)\right\}-\tilde\Delta_{22}-\lambda_2(2p)^{-1/2}, &(67)
\end{aligned}$$

where $\Gamma_1=\{(\gamma_0,\gamma_1^{\mathrm{T}},\gamma_2^{\mathrm{T}})^{\mathrm{T}}:\gamma_2=0\}$. Inequality (65) follows because Γ_1 is defined as a subset of Γ such that $\gamma_2=0$ and hence $\mathrm{pen}_2(\gamma_2)=0$ for $\gamma\in\Gamma_1$.

Take $\Gamma_{10}=\{\gamma\in\Gamma_{\mathrm{rp}1}:\gamma_0=0,\mathrm{pen}_2(\gamma)=(2p)^{-1/2}\}$, where $\Gamma_{\mathrm{rp}1}$ is the subset of Γ_1 associated with main-effect ramp functions as in the proof of Theorem 12. Inequality (66) holds because $\Gamma_{10}\subset\Gamma_1$ by definition. Inequality (67) follows from Proposition 55: it holds with probability at least $1-2\delta$ that for any $\gamma\in\Gamma_{10}$,

$$K_{\mathrm{HG}}(P_n,P_{\hat\theta};h_{\gamma,\hat\mu})\ge\mathrm{E}_{P_{\theta^*}}h_{\gamma,\hat\mu}(x)-\mathrm{E}_{P_{\hat\theta}}h_{\gamma,\hat\mu}(x)-\tilde\Delta_{22},$$

where $\tilde\Delta_{22}=\epsilon+\lambda_{22}(2p)^{-1/2}$, and

$$\lambda_{22}=C_{\mathrm{rad}5}\sqrt{\frac{16p}{n}}+\sqrt{\frac{2p\log(\delta^{-1})}{n}}.$$

Note that λ_{22} is the same as in the proof of Theorem 12. From (65)–(67), the lower bound in (62) holds with probability at least $1-2\delta$, where $\Delta_{22}=\tilde\Delta_{22}+\lambda_2(2p)^{-1/2}$ and $d(\hat\theta,\theta^*)=\mathrm{E}_{P_{\theta^*}}h_{\gamma,\hat\mu}(x)-\mathrm{E}_{P_{\hat\theta}}h_{\gamma,\hat\mu}(x)$.

(Step 3) We deduce upper bounds on $\|\hat{\mu}-\mu^*\|_2$ and $\|\hat{\sigma}-\sigma^*\|_2$, by choosing appropriate b_2 and relating the moment matching term $d(\hat{\theta},\theta^*)$ to the estimation errors. First, combining the upper bound in (62) from Step 1 and the lower bound from Step 2 shows that with probability at least $1-6\delta$,

$$\begin{aligned}&(2p)^{-1/2}\max_{\gamma\in\Gamma_{\mathrm{rp1}},\mathrm{pen}_2(\gamma)=1}\left\{\mathrm{E}_{P_{\theta^*}}h_{\gamma,\hat{\mu}}(x)-\mathrm{E}_{P_{\hat{\theta}}}h_{\gamma,\hat{\mu}}(x)\right\}\\&\leq 3\epsilon+2\sqrt{\epsilon/(n\delta)}+(\lambda_2+\lambda_{22})(2p)^{-1/2},\end{aligned}$$

which gives

$$\max_{\gamma\in\Gamma_{\mathrm{rp1}},\mathrm{pen}_2(\gamma)=\sqrt{1/2}}\left\{\mathrm{E}_{P_{\theta^*}}h_{\gamma,\hat{\mu}}(x)-\mathrm{E}_{P_{\hat{\theta}}}h_{\gamma,\hat{\mu}}(x)\right\}\leq \mathrm{Err}_{h2}(n,p,\delta,\epsilon), \tag{68}$$

where

$$\mathrm{Err}_{h2}(n,p,\delta,\epsilon)=3\epsilon\sqrt{p}+2\sqrt{p\epsilon/(n\delta)}+(\lambda_2+\lambda_{22})/\sqrt{2}.$$

The desired result then follows from Proposition 48: provided $\mathrm{Err}_{h2}(n,p,\delta,\epsilon)\leq a$, inequality (68) implies that

$$\begin{aligned}\|\hat{\mu}-\mu^*\|_2&\leq S_{4,a}\mathrm{Err}_{h2}(n,p,\delta,\epsilon),\\\|\hat{\sigma}-\sigma^*\|_2&\leq S_{6,a}\mathrm{Err}_{h2}(n,p,\delta,\epsilon).\end{aligned}$$

(Step 4) For the second version of the lower bound in (49), we show that with probability at least $1-2\delta$,

$$\begin{aligned}&\max_{\gamma\in\Gamma}\left\{K_{\mathrm{HG}}(P_n,P_{\hat{\theta}};h_{\gamma,\hat{\mu}})-\lambda_2\,\mathrm{pen}_2(\gamma_1)-\lambda_3\,\mathrm{pen}_2(\gamma_2)\right\}\\&\geq\max_{\gamma\in\Gamma_2}\left\{K_{\mathrm{HG}}(P_n,P_{\hat{\theta}};h_{\gamma,\hat{\mu}})-\lambda_3\,\mathrm{pen}_2(\gamma_2)\right\} &(69)\\&\geq\max_{\gamma\in\Gamma_{20}}\left\{K_{\mathrm{HG}}(P_n,P_{\hat{\theta}};h_{\gamma,\hat{\mu}})\right\}-\lambda_3\,(4q)^{-1/2} &(70)\\&\geq\max_{\gamma\in\Gamma_{20}}\left\{\mathrm{E}_{P_{\theta^*}}h_{\gamma,\hat{\mu}}(x)-\mathrm{E}_{P_{\hat{\theta}}}h_{\gamma,\hat{\mu}}(x)\right\}-\tilde{\Delta}_{32}-\lambda_3(4q)^{-1/2}, &(71)\end{aligned}$$

where $\Gamma_2=\{(\gamma_0,\gamma_1^{\mathrm{T}},\gamma_2^{\mathrm{T}})^{\mathrm{T}}:\gamma_1=0\}$. Inequality (69) follows because Γ_2 is a subset of Γ such that $\gamma_1=0$ and hence $\mathrm{pen}_2(\gamma_1)=0$ for $\gamma\in\Gamma_2$.

Take $\Gamma_{20}=\{\gamma\in\Gamma_{\mathrm{rp2}}:\Gamma_0=0,\mathrm{pen}_2(\gamma)=(4q)^{-1/2}\}$ for $q=p(1-p)$, where Γ_{rp2} is the subset of Γ_2 associated with interaction ramp functions as in the proof of Theorem 12. Inequality (70) holds because $\Gamma_{20}\subset\Gamma_2$ by definition. Inequality (71) follows from Proposition 56: it holds with probability at least $1-2\delta$ that for any $\gamma\in\Gamma_{20}$,

$$K_{\mathrm{HG}}(P_n,P_{\hat{\theta}};h_{\gamma,\hat{\mu}})\geq\mathrm{E}_{P_{\theta^*}}h(x)-\mathrm{E}_{P_{\hat{\theta}}}h(x)-\tilde{\Delta}_{32},$$

where $\tilde{\Delta}_{32}=\epsilon+\sqrt{p}\lambda_{32}(4q)^{-1/2}$ and

$$\lambda_{32}=C_{\mathrm{rad4}}\sqrt{\frac{6(p-1)}{n}}+\sqrt{\frac{(p-1)\log(\delta^{-1})}{n}}.$$

Note that λ_{32} is the same as in the proof of Theorem 12. From (69)–(71), the lower bound in (62) holds with probability at least $1-2\delta$, where $\Delta_{22} = \tilde{\Delta}_{32} + \lambda_3(4q)^{-1/2}$ and $d(\hat{\theta}, \theta^*) = \max_{\gamma\in\Gamma_{20}}\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\}$.

(Step 5) We deduce an upper bound on $\|\hat{\Sigma} - \Sigma^*\|_{\mathrm{F}}$, by relating the moment matching term $d(\hat{\theta}, \theta^*)$ to the estimation error. First, combining the upper bound in (62) from Step 1 and the lower bound from Step 4 shows that with probability $1-6\delta$,

$$\begin{aligned}&(4q)^{-1/2} \max_{\gamma\in\Gamma_{\mathrm{rp2}}, \mathrm{pen}_2(\gamma)=1} \left\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\right\} \\ &\le 3\epsilon + 2\sqrt{\epsilon/(n\delta)} - \sqrt{p}\lambda_{32}(4q)^{-1/2} - (25/3)\sqrt{5p}C_{\mathrm{sp}22}M_{21}\lambda_{31}(4q)^{-1/2},\end{aligned}$$

which gives

$$\max_{\gamma\in\Gamma_{\mathrm{rp2}}, \mathrm{pen}_2(\gamma)=1/2} \left\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\right\} \le \sqrt{p}\,\mathrm{Err}_{h3}(n,p,\delta,\epsilon), \tag{72}$$

where

$$\mathrm{Err}_{h3}(n,p,\delta,\epsilon) = 3\epsilon\sqrt{p-1} + 2\sqrt{\epsilon(p-1)/(n\delta)} + \lambda_{32}/2 + (25\sqrt{5}/6)C_{\mathrm{sp}22}M_{21}\lambda_{31}$$

The desired result then follows from Proposition 50: inequality (72) implies that

$$\begin{aligned}\frac{1}{\sqrt{p}}\|\hat{\Sigma} - \Sigma^*\|_{\mathrm{F}} &\le 2M_2^{1/2}\|\hat{\sigma} - \sigma^*\|_2 + S_7(\sqrt{2}\Delta_{\hat{\mu},\hat{\sigma}} + \mathrm{Err}_{h3}(n,p,\delta,\epsilon)) \\ &\le S_{9,a}\mathrm{Err}_{h2}(n,p,\delta,\epsilon) + S_7\mathrm{Err}_{h3}(n,p,\delta,\epsilon),\end{aligned}$$

where $\Delta_{\hat{\mu},\hat{\sigma}} = (\|\hat{\mu} - \mu^*\|_2^2 + \|\hat{\sigma} - \sigma^*\|_2^2)^{1/2}$ and $S_{9,a} = 2M_2^{1/2}S_{6,a} + \sqrt{2}S_7(S_{4,a} + S_{6,a})$. ■

B.4 Proof of Corollary 18

(i) In the proofs of Theorems 11 and 12, we used the main frame,

$$d(\hat{\theta}, \theta^*) - \Delta_1 \le \max_{\gamma\in\Gamma}\{K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \mathrm{pen}(\gamma;\lambda)\} \le \Delta_2, \tag{73}$$

where $\mathrm{pen}(\gamma;\lambda)$ is $\lambda_1(\|\gamma_1\|_1 + \|\gamma_2\|_1)$ or $\lambda_2\|\gamma_1\|_2 + \lambda_3\|\gamma_2\|_2$. For Theorems 11 and 12, we showed the upper bound in (73) using the fact that $\hat{\theta}$ is the minimizer of

$$\max_{\gamma\in\Gamma}\{K_f(P_n, P_\theta; h_{\gamma,\mu}) - \lambda\mathrm{pen}(\gamma)\},$$

which is a function of θ by the definition of (17) and (19) as nested optimization (see Remark 1). Now $\hat{\theta}$ is not defined as a minimizer of the above function, but a solution to an alternating optimization problem (22) with two objectives. We need to develop new arguments. On the other hand, we showed the lower bound in (73) for Theorems 11 and 12, through choosing different subsets of Γ. The previous arguments are still applicable here.

(Step 1) For the upper bound in (73), we show that the following holds with probability at least $1-\delta$,

$$\begin{aligned}
&\max_{\gamma\in\Gamma}\{K_f(P_n, P_{\hat\theta}; h_{\gamma,\hat\mu}) - \mathrm{pen}(\gamma;\lambda)\} \\
&\le \max_{\gamma\in\Gamma}\{-f'(1-\hat\epsilon)\hat\epsilon + R_1\left|\mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\hat\mu}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x)\right| - \mathrm{pen}(\gamma;\lambda)\} \qquad (74)\\
&\le \Delta_1 + \max_{\gamma\in\Gamma}\{R_1\left|\mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\hat\mu}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x)\right| - \mathrm{pen}(\gamma;\lambda)\}, \qquad (75)
\end{aligned}$$

where $\hat\epsilon$ is the (unobserved) fraction of contamination in $(X_1,\ldots,X_n)$. Inequality (74) follows from Lemma 57, and is the most important step for connecting two-objective GAN with logit f-GAN. Inequality (75) follows from an upper bound on $\hat\epsilon$ as proved in Proposition 33, where $\Delta_1 = -f'(3/5)(\epsilon+\sqrt{\epsilon/(n\delta)})$, the same as Δ_{11} and Δ_{21} in the proofs of Theorems 11 and 12.

Similarly as in Proposition 33 or 44, the term $|\mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\hat\mu}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x)|$ can be controlled in terms of the L_1 or L_2 norms of (γ_1,γ_2), using Lemma 30 or 43. Then for $\mathrm{pen}(\gamma;\lambda)$ defined as an L_1 or L_2 penalty, it can be shown that the following holds with probability at least $1-4\delta$ or $1-6\delta$,

$$R_1\left|\mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\hat\mu}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat\mu}(x)\right| - \mathrm{pen}(\gamma;\lambda) \le 0. \qquad (76)$$

provided that the tuning parameters λ_1 or (λ_2,λ_3) are chosen as in Theorem 11 or 12 respectively. From (74)–(76), the upper bound holds in (73) with probability $1-5\delta$ or $1-7\delta$.

(Steps 2,3) The lower bound step and the estimation error step for L_1 or L_2 penalized two-objective GAN are the same as in the proofs of Theorems 11 and 12 respectively.

(ii) For the two-objective hinge GAN hinge (23), the result follows similarly using Lemma 58 with $\Delta_1 = 2(\epsilon+\sqrt{\epsilon/(n\delta)})$ and $R_1 = 1$.

Appendix C. Technical details

C.1 Details in main proof of Theorem 6

Lemma 27 *Suppose that $f:(0,\infty)\to\mathbb{R}$ is convex with $f(1)=0$ and satisfies Assumption 1(i). Denote $C_f = \inf_{t\in(0,1]} f''(t)$. Then*

$$D_f(P||Q) \ge \frac{C_f}{2}\mathrm{TV}(P,Q)^2.$$

If further Assumption 1(iii) holds, then $D_f(P||Q) \ge \frac{f''(1)}{2}\mathrm{TV}(P,Q)^2$.

Proof Because x^2 is convex in x, by Jensen's inequality, we have

$$\mathrm{TV}(P,Q)^2 \le \int f_{\mathrm{TV}}^2(p/q)\mathrm{d}Q,$$

where $f_{\mathrm{TV}}(t) = (1-t)_+$ and p/q is the density ratio $\mathrm{d}P/\mathrm{d}Q$. Note that $D_f(P\|Q)$ can be equivalently obtained as $D_{\tilde f}(P\|Q)$, where $\tilde f(t) = f(t) - f'(1)(t-1)$. Therefore, it suffices to show that $\tilde f(t) \ge \frac{C_f}{2} f_{TV}^2(t)$ for $t\in(0,\infty)$.

By a Taylor expansion of f, we have

$$f(t) = f(1) + f'(1)(t-1) + \frac{f''(\tilde{t})}{2}(t-1)^2$$
$$\geq f'(1)(t-1) + \frac{C_f}{2}(t-1)^2, \tag{77}$$

where $\tilde{t}$ lies between t and 1. If $t \in (0,1]$, then (77) gives

$$\tilde{f}(t) \geq \frac{C_f}{2}(t-1)^2 = \frac{C_f}{2}(1-t)_+^2 = \frac{C_f}{2} f_{\mathrm{TV}}^2(t).$$

If $t \in (1,\infty)$, then because $C_f \geq 0$ by convexity of f, (77) gives

$$\tilde{f}(t) \geq \frac{C_f}{2}(t-1)^2 \geq 0 = \frac{C_f}{2}(1-t)_+^2 = \frac{C_f}{2} f_{TV}^2(t).$$

Combining the two cases completes the proof. ■

Denote as $\Phi(\cdot)$ the cumulative distribution function of $\mathrm{N}(0,1)$, and $\mathrm{erf}(x)$ the probability of $[-\sqrt{2}x, \sqrt{2}x]$ under $\mathrm{N}(0,1)$ for $x \geq 0$.

Lemma 28 *Let $a \in [0, 1/2)$ be arbitrarily fixed.*

(i) If $\Phi(x) \leq 1/2 + a$ for $x \geq 0$, then

$$x \leq S_{1,a} \left\{\Phi(x) - 1/2\right\},$$

where $S_{1,a} = \{\Phi'(\Phi^{-1}(1/2+a))\}^{-1}$.

(ii) If $|\mathrm{erf}(x\sqrt{z_0/2}) - 1/2| \leq a$ for $x \geq 0$, then

$$|x-1| \leq S_{2,a} \left|\mathrm{erf}(x\sqrt{z_0/2}) - 1/2\right|,$$

where $S_{2,a} = \{\sqrt{z_0/2}\mathrm{erf}'(\sqrt{2/z_0}\mathrm{erf}^{-1}(1/2+a))\}^{-1}$ and z_0 is an universal constant such that $\mathrm{erf}(\sqrt{z_0/2}) = 1/2$.

Proof (i) By the mean value theorem, we have $\Phi(x) \geq \frac{1}{2} + S_{1,a}^{-1} x$, because $\Phi'(\cdot)$ is decreasing on $[0, +\infty)$.

(ii) By the mean value theorem, we have $|\mathrm{erf}(x\sqrt{z_0/2}) - 1/2| \geq S_{2,a}^{-1}|x-1|$, because $\mathrm{erf}'(\cdot)$ is decreasing on $[0, +\infty)$. ■

Proposition 29 *For two multivariate Gaussian distributions, $P_{\bar{\theta}}$ and P_{θ^*}, with $\bar{\theta} = (\bar{\mu}, \bar{\Sigma})$ and $\theta^* = (\mu^*, \Sigma^*)$, denote $d(\bar{\theta}, \theta^*) = \mathrm{TV}(P_{\bar{\theta}}, P_{\theta^*})$.*

(i) If $d(\bar{\theta}, \theta^*) \leq a$ for a constant $a \in [0, 1/2)$, then

$$\|\bar{\mu} - \mu^*\|_2 \leq S_{1,a} \|\Sigma^*\|_{\mathrm{op}}^{1/2} d(\bar{\theta}, \theta^*),$$
$$\|\bar{\mu} - \mu^*\|_\infty \leq S_{1,a} \|\Sigma^*\|_{\max}^{1/2} d(\bar{\theta}, \theta^*),$$

where $S_{1,a} = \{\Phi'(\Phi^{-1}(1/2+a))\}^{-1}$ as in Lemma 28.

(ii) If further $d(\bar{\theta}, \theta^*) \le a/(1 + S_{1,a})$*, then*

$$\|\bar{\Sigma} - \Sigma^*\|_{\mathrm{op}} \le 2S_{3,a}\|\Sigma^*\|_{\mathrm{op}} d(\bar{\theta}, \theta^*) + S_{3,a}^2 \|\Sigma^*\|_{\mathrm{op}} (d(\bar{\theta}, \theta^*))^2,$$
$$\|\bar{\Sigma} - \Sigma^*\|_{\max} \le 4S_{3,a}\|\Sigma^*\|_{\max} d(\bar{\theta}, \theta^*) + 2S_{3,a}^2 \|\Sigma^*\|_{\max} (d(\bar{\theta}, \theta^*))^2,$$

where $S_{3,a} = S_{2,a}(1 + S_{1,a})$*,* $S_{2,a} = \{\sqrt{z_0/2}\,\mathrm{erf}'(\sqrt{2/z_0}\,\mathrm{erf}^{-1}(1/2 + a))\}^{-1}$*, and the constant* z_0 *is defined such that* $\mathrm{erf}(\sqrt{z_0/2}) = 1/2$*, as in Lemma 28.*

Proof The TV distance, $D_{\mathrm{TV}}(P_1 \| P_2)$, can be equivalently defined as

$$\mathrm{TV}(P_1, P_2) = \sup_{A \in \mathcal{A}} |P_1(A) - P_2(A)|,$$

for P_1 and P_2 defined in a probability space $(\mathcal{X}, \mathcal{A})$. This definition is applicable to multivariate Gaussian distributions with singular variance matrices. To derive the desired results, we choose specific events A and show that the differences in the means and variance matrices can be upper bounded by $|P_{\bar{\theta}}(A) - P_{\theta^*}(A)|$.

We first show results (i) and (ii), when Σ^* and $\bar{\Sigma}$ are nonsingular. Then we show that the results remain valid when Σ^* or $\bar{\Sigma}$ is singular.

(i) Assume that both Σ^* and $\bar{\Sigma}$ are nonsingular. For any $u \in \mathbb{R}^p$, we have by the definition of TV,

$$P_{\mu^*, \Sigma^*}(u^{\mathrm{T}} X \le u^{\mathrm{T}} \bar{\mu}) - P_{\bar{\mu}, \bar{\Sigma}}(u^{\mathrm{T}} X \le u^{\mathrm{T}} \bar{\mu}) \le d(\bar{\theta}, \theta^*).$$

For nonzero $u \in \mathbb{R}^p$, because $u^{\mathrm{T}} \bar{\Sigma} u \ne 0$ and $u^{\mathrm{T}} \Sigma^* u \ne 0$, we have

$$P_{\bar{\mu}, \bar{\Sigma}}(u^{\mathrm{T}} X \le u^{\mathrm{T}} \bar{\mu}) = \frac{1}{2},$$
$$P_{\mu^*, \Sigma^*}(u^{\mathrm{T}} X \le u^{\mathrm{T}} \bar{\mu}) = \Phi\left(\frac{u^{\mathrm{T}}(\bar{\mu} - \mu^*)}{\sqrt{u^{\mathrm{T}} \Sigma^* u}}\right).$$

Combining the preceding three displays shows that for nonzero $u \in \mathbb{R}^p$,

$$\Phi\left(\frac{u^{\mathrm{T}}(\bar{\mu} - \mu^*)}{\sqrt{u^{\mathrm{T}} \Sigma^* u}}\right) \le \frac{1}{2} + d(\bar{\theta}, \theta^*).$$

By Lemma 28 (i), if $d(\bar{\theta}, \theta^*) \le a$ for a constant $a \in [0, 1/2)$, then for any $u \in \mathbb{R}^p$ satisfying $u^{\mathrm{T}}(\bar{\mu} - \mu^*) \ge 0$,

$$0 \le u^{\mathrm{T}}(\bar{\mu} - \mu^*) \le \sqrt{u^{\mathrm{T}} \Sigma^* u} S_{1,a} d(\bar{\theta}, \theta^*). \tag{78}$$

Let $\mathcal{U}_2 = \{u \in \mathbb{R}^p : \|u\|_2 = 1\}$. By (78) with u restricted such that $u \in \mathcal{U}_2$ and $u^{\mathrm{T}}(\bar{\mu} - \mu^*) \ge 0$, we have

$$\|\bar{\mu} - \mu^*\|_2 = \sup_{u \in \mathcal{U}_2} u^{\mathrm{T}}(\bar{\mu} - \mu^*)$$
$$\le S_{1,a} \|\Sigma^*\|_{\mathrm{op}}^{1/2} d(\bar{\theta}, \theta^*).$$

Similarly, let $\mathcal{U}_\infty = \{\pm \mathrm{e}_j : j = 1, \ldots, p\}$, where e_j is a vector with jth coordinate being one and others being zero. By (78) with u restricted such that $u \in \mathcal{U}_\infty$ and $u^{\mathrm{T}}(\bar{\mu} - \mu^*) \geq 0$, we have

$$\begin{aligned}\|\bar{\mu} - \mu^*\|_\infty &= \sup_{u \in \mathcal{U}_\infty} u^{\mathrm{T}}(\bar{\mu} - \mu^*) \\ &\leq S_{1,a} \|\Sigma^*\|_{\max}^{1/2} d(\bar{\theta}, \theta^*).\end{aligned}$$

The last line uses the fact that $\sup_{u \in \mathcal{U}_\infty} u^{\mathrm{T}} \Sigma^* u = \|\mathrm{diag}(\Sigma^*)\|_\infty = \|\Sigma^*\|_{\max}$ by the nature of variance matrices.

(ii) Assume that Σ^* and $\bar{\Sigma}$ are nonsingular. We first separate the bias caused by the location difference between $P_{\bar{\theta}}$ and P_{θ^*}. By the triangle inequality, we have

$$\mathrm{TV}(P_{\bar{\mu},\Sigma^*}, P_{\bar{\mu},\bar{\Sigma}}) \leq \mathrm{TV}(P_{\bar{\mu},\Sigma^*}, P_{\mu^*,\Sigma^*}) + \mathrm{TV}(P_{\mu^*,\Sigma^*}, P_{\bar{\mu},\bar{\Sigma}}). \tag{79}$$

By Lemma 27, we know that $TV(P, Q) \leq \sqrt{2 D_{\mathrm{KL}}(P||Q)}$. Then we have

$$\begin{aligned}\mathrm{TV}(P_{\bar{\mu},\Sigma^*}, P_{\mu^*,\Sigma^*}) &\leq \sqrt{2 D_{\mathrm{KL}}(P_{\bar{\mu},\Sigma^*}||P_{\mu^*,\Sigma^*})} \\ &\leq S_{1,a} d(\bar{\theta}, \theta^*). \end{aligned} \tag{80}$$

provided that $d(\bar{\theta}, \theta^*) \leq a$. Inequality (80) follows because by standard calculation

$$D_{\mathrm{KL}}(N(\bar{\mu}, \Sigma^*)||N(\mu^*, \Sigma^*)) = \frac{1}{2}(\bar{\mu} - \mu^*)^{\mathrm{T}} \Sigma^{*-1} (\bar{\mu} - \mu^*),$$

and taking $u = \Sigma^{*-1}(\bar{\mu} - \mu^*)$ in (78) gives

$$\sqrt{(\bar{\mu} - \mu^*)^{\mathrm{T}} \Sigma^{*-1} (\bar{\mu} - \mu^*)} \leq S_{1,a} d(\bar{\theta}, \theta^*).$$

Combining (79) and (80) yields

$$\mathrm{TV}(P_{\bar{\mu},\Sigma^*}, P_{\bar{\mu},\bar{\Sigma}}) \leq d(\bar{\theta}, \theta^*) + S_{1,a} d(\bar{\theta}, \theta^*). \tag{81}$$

For any $u \in \mathbb{R}^p$ such that $u^{\mathrm{T}}(\bar{\Sigma} - \Sigma^*)u \geq 0$, (81) implies

$$\begin{aligned}0 &\leq P_{\bar{\mu},\Sigma^*}\left\{(u^{\mathrm{T}}X - u^{\mathrm{T}}\bar{\mu})^2 \leq z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} - P_{\bar{\mu},\bar{\Sigma}}\left\{(u^{\mathrm{T}}X - u^{\mathrm{T}}\bar{\mu})^2 \leq z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} \\ &= P_{0,\Sigma^*}\left\{(u^{\mathrm{T}}X)^2 \leq z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} - P_{0,\bar{\Sigma}}\left\{(u^{\mathrm{T}}X)^2 \leq z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} \\ &\leq d(\bar{\theta}, \theta^*) + S_{1,a} d(\bar{\theta}, \theta^*),\end{aligned}$$

where z_0 is an universal constant such that $\mathrm{erf}(\sqrt{z_0/2}) = 1/2$. Similarly, for any $u \in \mathbb{R}^p$ such that $u^{\mathrm{T}}(\bar{\Sigma} - \Sigma^*)u \leq 0$, (81) implies

$$\begin{aligned}0 &\leq P_{\bar{\mu},\Sigma^*}\left\{(u^{\mathrm{T}}X - u^{\mathrm{T}}\bar{\mu})^2 \geq z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} - P_{\bar{\mu},\bar{\Sigma}}\left\{(u^{\mathrm{T}}X - u^{\mathrm{T}}\bar{\mu})^2 \geq z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} \\ &= P_{0,\Sigma^*}\left\{(u^{\mathrm{T}}X)^2 \geq z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} - P_{0,\bar{\Sigma}}\left\{(u^{\mathrm{T}}X)^2 \geq z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} \\ &\leq d(\bar{\theta}, \theta^*) + S_{1,a} d(\bar{\theta}, \theta^*).\end{aligned}$$

Notice that the choice of z_0 ensures that for $Z \in \mathrm{N}(0,1)$,

$$\begin{aligned} P_{0,\bar{\Sigma}}\left\{(u^{\mathrm{T}}X)^2 \le z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} &= \mathrm{P}(Z^2 \le z_0) = \frac{1}{2} \\ &= \mathrm{P}(Z^2 \ge z_0) = P_{0,\bar{\Sigma}}\left\{(u^{\mathrm{T}}X)^2 \ge z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\}. \end{aligned}$$

Moreover, for any nonzero $u \in \mathbb{R}^p$, we have by the definition of erf,

$$P_{0,\Sigma^*}\left\{(u^{\mathrm{T}}X)^2 \le z_0 u^{\mathrm{T}}\bar{\Sigma}u\right\} = \mathrm{erf}\left(\sqrt{\frac{z_0 u^{\mathrm{T}}\bar{\Sigma}u}{2u^{\mathrm{T}}\Sigma^* u}}\right).$$

Combining the preceding four displays shows that for any nonzero $u \in \mathbb{R}^p$,

$$\left|\mathrm{erf}\left(\sqrt{\frac{z_0 u^{\mathrm{T}}\bar{\Sigma}u}{2u^{\mathrm{T}}\Sigma^* u}}\right) - \frac{1}{2}\right| \le (1 + S_{1,a})d(\bar{\theta}, \theta^*).$$

By Lemma 28 (ii), if $d(\bar{\theta}, \theta^*) \le \min(a, a/(1+S_{1,a})) = a/(1+S_{1,a})$, then for any nonzero $u \in \mathbb{R}^p$,

$$\left|\sqrt{\frac{u^{\mathrm{T}}\bar{\Sigma}u}{u^{\mathrm{T}}\Sigma^* u}} - 1\right| \le S_{2,a}(1+S_{1,a})d(\bar{\theta}, \theta^*),$$

or equivalently for any $u \in \mathbb{R}^p$,

$$\left|\sqrt{u^{\mathrm{T}}\bar{\Sigma}u} - \sqrt{u^{\mathrm{T}}\Sigma^* u}\right| \le S_{2,a}(1+S_{1,a})\sqrt{u^{\mathrm{T}}\Sigma^* u}d(\bar{\theta}, \theta^*). \tag{82}$$

Notice that for any $a, b, c \ge 0$, if $|\sqrt{a} - \sqrt{b}| \le c$ then $|a - b| \le 2\sqrt{b}c + c^2$. Thus, inequality (82) implies

$$\begin{aligned} \|\bar{\Sigma} - \Sigma^*\|_{\mathrm{op}} &= \sup_{u \in \mathcal{U}_2} \left|u^{\mathrm{T}}(\bar{\Sigma} - \Sigma^*)u\right| \\ &\le 2S_{3,a}\|\Sigma^*\|_{\mathrm{op}}d(\bar{\theta}, \theta^*) + S_{3,a}^2\|\Sigma^*\|_{\mathrm{op}}(d(\bar{\theta}, \theta^*))^2, \end{aligned} \tag{83}$$

where $S_{3,a} = S_{2,a}(1 + S_{1,a})$.

To handle $\|\bar{\Sigma} - \Sigma^*\|_{\max}$, let $\mathcal{U}_{2,\infty} = \{\pm e_{ij} : i, j = 1, \ldots, p, i \ne j\}$, where e_{ij} is a vector in $\mathcal{U}_2$ with only ith and jth coordinates possibly being nonzero. For $u \in \mathcal{U}_{2,\infty}$, we have $u^{\mathrm{T}}\Sigma^* u = u_{ij}^{\mathrm{T}}\Sigma_{ij}^* u_{ij}$ and $u^{\mathrm{T}}\bar{\Sigma}u = u_{ij}^{\mathrm{T}}\bar{\Sigma}_{ij}u_{ij}$, where $u_{ij} \in \mathbb{R}^2$ is formed by ith and jth coordinates of u, and Σ_{ij}^* and $\bar{\Sigma}_{ij}$ are 2×2 matrices, formed by selecting ith and jth rows and columns from Σ^* and $\bar{\Sigma}$ respectively. Similarly as in the deviation of (83), applying inequality (82) with $u \in \mathcal{U}_{2,\infty}$, we have

$$\begin{aligned} \|\bar{\Sigma}_{ij} - \Sigma_{ij}^*\|_{\mathrm{op}} &= \sup_{u \in U_{2,\infty}} \left|u^{\mathrm{T}}(\bar{\Sigma} - \Sigma^*)u\right| \\ &\le 2S_{3,a}\|\Sigma_{ij}^*\|_{\mathrm{op}}d(\bar{\theta}, \theta^*) + S_{3,a}^2\|\Sigma_{ij}^*\|_{\mathrm{op}}(d(\bar{\theta}, \theta^*))^2. \end{aligned}$$

Because for a matrix $A \in \mathbb{R}^{m_1 \times m_2}$, $\|A\|_{\max} \le \|A\|_{\mathrm{op}} \le \sqrt{m_1 m_2}\|A\|_{\max}$, the above inequality implies that for any $i \ne j \in \{1, \ldots, p\}$,

$$\begin{aligned} &\|\bar{\Sigma}_{ij} - \Sigma_{ij}^*\|_{\max} \\ &\le 4S_{3,a}\|\Sigma_{ij}^*\|_{\max}d(\bar{\theta}, \theta^*) + 2S_{3,a}^2\|\Sigma_{ij}^*\|_{\max}(d(\bar{\theta}, \theta^*))^2. \end{aligned} \tag{84}$$

Taking the maximum on both sides of (84) over $i \neq j$ gives the desired result:

$$\begin{aligned}\|\bar{\Sigma} - \Sigma^*\|_{\max} &= \max_{i \neq j \in \{1,\ldots,p\}} \|\bar{\Sigma}_{ij} - \Sigma^*_{ij}\|_{\max} \\ &\leq 4S_{3,a}\|\Sigma^*\|_{\max} d(\bar{\theta}, \theta^*) + 2S_{3,a}^2 \|\Sigma^*\|_{\max} (d(\bar{\theta}, \theta^*))^2.\end{aligned}$$

(iii) Consider the case where Σ^* or $\bar{\Sigma}$ is singular. As the following argument is symmetric in Σ^* and $\bar{\Sigma}$, we assume without loss of generality that Σ^* is singular. Fix any nonzero u such that $u^{\mathrm{T}}\Sigma^* u = 0$.

First, we show that for $\bar{\theta} = (\bar{\mu}, \bar{\Sigma})$ such that $\mathrm{TV}(P_{\bar{\theta}}, P_{\theta^*}) < 1$, we also have $u^{\mathrm{T}}\bar{\Sigma}u = 0$. In fact, $\mathrm{TV}(P_{\bar{\theta}}, P_{\theta^*}) < 1$ implies

$$\left|P_{\mu^*,\Sigma^*}(u^{\mathrm{T}}X = u^{\mathrm{T}}\mu^*) - P_{\bar{\mu},\bar{\Sigma}}(u^{\mathrm{T}}X = u^{\mathrm{T}}\mu^*)\right| \leq d(\bar{\theta}, \theta^*) < 1. \tag{85}$$

Note that $P_{\mu^*,\Sigma^*}(u^{\mathrm{T}}X = u^{\mathrm{T}}\mu^*) = 1$ because $u^{\mathrm{T}}\Sigma^* u = 0$. If $u^{\mathrm{T}}\bar{\Sigma}u > 0$, then $P_{\bar{\mu},\bar{\Sigma}}(u^{\mathrm{T}}X = u^{\mathrm{T}}\mu^*) = 0$, and hence (85) gives

$$|1 - 0| \leq d(\bar{\theta}, \theta^*) < 1,$$

which is a contradiction. Thus $u^{\mathrm{T}}\bar{\Sigma}u = 0$.

Next we show that for $\bar{\theta} = (\bar{\mu}, \bar{\Sigma})$ such that $\mathrm{TV}(\bar{\theta}, \theta^*) < 1$, we also have $u^{\mathrm{T}}(\mu^* - \bar{\mu}) = 0$. In fact, with $u^{\mathrm{T}}\bar{\Sigma}u = 0$ as shown above, we have that $P_{\bar{\mu},\bar{\Sigma}}(u^{\mathrm{T}}X = u^{\mathrm{T}}\mu^*) = 1$ if $u^{\mathrm{T}}\mu^* = u^{\mathrm{T}}\bar{\mu}$ and $P_{\bar{\mu},\bar{\Sigma}}(u^{\mathrm{T}}X = u^{\mathrm{T}}\mu^*) = 0$ otherwise. If $u^{\mathrm{T}}\mu^* \neq u^{\mathrm{T}}\bar{\mu}$, then inequality (85) gives

$$|1 - 0| \leq d(\bar{\theta}, \theta^*) < 1,$$

which is a contradiction. Thus $u^{\mathrm{T}}\mu^* = u^{\mathrm{T}}\bar{\mu}$.

From the two preceding results, we see that the upper bounds (78) and (82) derived in (i) and (ii) remain valid for any $u \in \mathbb{R}^p$ satisfying $u^{\mathrm{T}}\Sigma^* u = 0$. Hence the desired results hold by the remaining proofs in (i) and (ii). ■

C.2 Details in main proof of Theorem 11

Lemma 30 *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed as $X \sim N_p(0, \Sigma)$ with $\|\Sigma\|_{\max} \leq M_1$. For k fixed knots $\xi_1, \ldots, \xi_k$ in $\mathbb{R}$, denote $\varphi(x) = (\varphi_1^{\mathrm{T}}(x), \ldots, \varphi_k^{\mathrm{T}}(x))^{\mathrm{T}}$, where $\varphi_l(x) \in \mathbb{R}^p$ is obtained by applying $t \mapsto (t - \xi_l)_+$ componentwise to $x \in \mathbb{R}^p$ for $l = 1, \ldots, k$. Then the following results hold.*

(i) Each component of the random vector $\varphi(X) - \mathrm{E}\varphi(X)$ is a sub-gaussian random variable with tail parameter $M_1^{1/2}$.

(ii) For any $\delta > 0$, we have that with probability at least $1 - 2\delta$,

$$\begin{aligned}&\sup_{\|w\|_1 = 1} \left| w^{\mathrm{T}} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi(X_i) - \mathrm{E}\varphi(X) \right\} \right| \\ &\leq C_{\mathrm{sp}11} M_1^{1/2} \sqrt{\frac{2\log(kp) + \log(\delta^{-1})}{n}},\end{aligned}$$

where $C_{\mathrm{sp11}} = \sqrt{2}C_{\mathrm{sg5}}$*, depending on the universal constant* C_{sg5} *in Lemma 69.*

(iii) Let $\mathcal{A}_1 = \{A \in \mathbb{R}^{kp\times kp} : \|A\|_{1,1} = 1\}$*. For any* $\delta > 0$*, we have that with probability at least* $1 - 4\delta$*, both the inequality in (ii) and*

$$\sup_{A\in\mathcal{A}_1} \left| \frac{1}{n}\sum_{i=1}^{n} \varphi^{\mathrm{T}}(X_i)A\varphi(X_i) - \mathrm{E}\varphi^{\mathrm{T}}(X)A\varphi(X) \right| \\ \leq C_{\mathrm{sp12}} M_{11} \left\{ \sqrt{\frac{2\log(kp) + \log(\delta^{-1})}{n}} + \frac{2\log(kp) + \log(\delta^{-1})}{n} \right\},$$

where $M_{11} = M_1^{1/2}(M_1^{1/2} + \sqrt{2\pi}\|\xi\|_\infty)$*,* $\|\xi\|_\infty = \max_{l=1,\ldots,k} |\xi_l|$*, and* $C_{\mathrm{sp12}} = \sqrt{2/\pi}C_{\mathrm{sp11}} + C_{\mathrm{sx7}}C_{\mathrm{sx6}}C_{\mathrm{sx5}}$*. Constants* $(C_{\mathrm{sx5}}, C_{\mathrm{sx6}}, C_{\mathrm{sx7}})$ *are the universal constants in Lemmas 74, 75, and 76.*

Proof (i) This can be obtained as the univariate case of Lemma 43 (i). It is only required that the marginal variance of each component of X is upper bounded by M_1.

(ii) Notice that

$$\sup_{\|w\|_1=1} \left| w^{\mathrm{T}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) - \mathrm{E}\varphi(X) \right\} \right| = \left\| \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) - \mathrm{E}\varphi(X) \right\|_\infty .$$

By (i) and sub-gaussian concentration (Lemma 69), each component of $n^{-1}\sum_{i=1}^{n}\varphi(X_i) - \mathrm{E}\varphi(X)$ is sub-gaussian with tail parameter $C_{\mathrm{sg5}}(M_1/n)^{1/2}$. Then for any $t > 0$, by the union bound, we have that with probability at $1 - 2k^2p^2\mathrm{e}^{-t}$,

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) - \mathrm{E}\varphi(X) \right\|_\infty \leq \sqrt{2}C_{\mathrm{sg5}}(M_1/n)^{1/2}t^{1/2}.$$

Taking $t = 2\log(kp) + \log(\delta^{-1})$ gives the desired result.

(iii) The difference of interest can be expressed in terms of the centered variables as

$$\frac{1}{n}\sum_{i=1}^{n} \varphi_i^{\mathrm{T}}A\varphi_i - \mathrm{E}\varphi^{\mathrm{T}}A\varphi \\ = \frac{1}{n}\sum_{i=1}^{n} (\varphi_i - \mathrm{E}\varphi)^{\mathrm{T}}A(\varphi_i - \mathrm{E}\varphi) - \mathrm{E}\{(\varphi - \mathrm{E}\varphi)^{\mathrm{T}}A(\varphi - \mathrm{E}\varphi)\} \quad (86) \\ + \frac{1}{n}\sum_{i=1}^{n} 2(\mathrm{E}\varphi)^{\mathrm{T}}A(\varphi_i - \mathrm{E}\varphi). \quad (87)$$

We handle the concentration of the two terms separately. Denote $\varphi_i = \varphi(X_i)$, $\varphi = \varphi(X)$, $\tilde{\varphi}_i = \varphi_i - \mathrm{E}\varphi$, and $\tilde{\varphi} = \varphi - \mathrm{E}\varphi$.

First, for $A \in \mathcal{A}_1$, the term in (87) can be bounded as follows:

$$\left| 2(\mathrm{E}\varphi)^{\mathrm{T}}A\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i \right| \leq 2\|\mathrm{E}\varphi\|_\infty \|A\|_{1,1} \left\| \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i \right\|_\infty = 2\|\mathrm{E}\varphi\|_\infty \left\| \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i \right\|_\infty \\ \leq 2\left(\frac{M_1^{1/2}}{\sqrt{2\pi}} + \|\xi\|_\infty \right) \left\| \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i \right\|_\infty ,$$

where $\|\xi\|_\infty = \max_{l=1,\ldots,k} |\xi_l|$. The second step holds because $\|A\|_{1,1} = 1$ for $A \in \mathcal{A}_1$. The third step holds because $\|\mathrm{E}\varphi_l\|_\infty \le M_1^{1/2}/\sqrt{2\pi} + |\xi_l|$ for $l = 1, \ldots, 5$ by Lemma 59. By (ii), for any $\delta > 0$, we have that with probability at least $1 - 2\delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^n \tilde{\varphi}_i \right\|_\infty \le C_{\mathrm{sp11}} M_1^{1/2} \sqrt{\frac{2\log(kp) + \log(\delta^{-1})}{n}}.$$

From the preceding two displays, we obtain that with probability at least $1 - 2\delta$,

$$\begin{aligned} &\sup_{A \in \mathcal{A}_1} \left| 2(\mathrm{E}\varphi)^{\mathrm{T}} A \frac{1}{n} \sum_{i=1}^n \tilde{\varphi}_i \right| \\ &\le \sqrt{\frac{2}{\pi}} C_{\mathrm{sp11}} M_1^{1/2} \left(M_1^{1/2} + \sqrt{2\pi} \|\xi\|_\infty \right) \sqrt{\frac{2\log(kp) + \log(\delta^{-1})}{n}}. \end{aligned} \tag{88}$$

Next, notice that

$$\sup_{A \in \mathcal{A}_1} \left| \frac{1}{n} \sum_{i=1}^n \tilde{\varphi}_i^{\mathrm{T}} A \tilde{\varphi}_i - \mathrm{E}\tilde{\varphi}^{\mathrm{T}} A \tilde{\varphi} \right| = \left\| \frac{1}{n} \sum_{i=1}^n \tilde{\varphi}_i \otimes \tilde{\varphi}_i - \mathrm{E}\tilde{\varphi} \otimes \tilde{\varphi} \right\|_{\max}.$$

From (i), each component of $\tilde{\varphi}_i$ is sub-gaussian with tail parameter $M_1^{1/2}$. By Lemma 74, each element of $\tilde{\varphi}_i \otimes \tilde{\varphi}_i$ is sub-exponential with tail parameter $C_{\mathrm{sx5}} M_1$. By Lemma 75, each element of the centered version, $\tilde{\varphi}_i \otimes \tilde{\varphi}_i - \mathrm{E}\tilde{\varphi} \otimes \tilde{\varphi}$, is sub-exponential with tail parameter $C_{\mathrm{sx6}} C_{\mathrm{sx5}} M_1$. Then for any $t > 0$, by Lemma 76 and the union bound, we have that with probability at least $1 - 2k^2 p^2 \mathrm{e}^{-t}$,

$$\left\| \frac{1}{n} \sum_{i=1}^n \tilde{\varphi}_i \otimes \tilde{\varphi}_i - \mathrm{E}\tilde{\varphi} \otimes \tilde{\varphi} \right\|_{\max} \le C_{\mathrm{sx7}} C_{\mathrm{sx6}} C_{\mathrm{sx5}} M_1 \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \right).$$

Taking $t = 2\log(kp) + \log(\delta^{-1})$, we obtain that with probability at least $1 - 2\delta$,

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^n \tilde{\varphi}_i \otimes \tilde{\varphi}_i - \mathrm{E}\tilde{\varphi} \otimes \tilde{\varphi} \right\|_{\max} \\ &\le C_{\mathrm{sx7}} C_{\mathrm{sx6}} C_{\mathrm{sx5}} M_1 \left\{ \sqrt{\frac{2\log(kp) + \log(\delta^{-1})}{n}} \vee \frac{2\log(kp) + \log(\delta^{-1})}{n} \right\}. \end{aligned} \tag{89}$$

Combining the two bounds (88) and (89) gives the desired result. ■

Lemma 31 *Suppose that $f : (0, \infty) \to \mathbb{R}$ is convex, non-increasing, and differentiable, and $f(1) = 0$. Denote $f^{\#}(t) = tf'(t) - f(t)$.*

(i) For any $t > 0$ and $\epsilon \in [0, 1)$, we have

$$(1 - \epsilon) f'(t) - f^{\#}(t) \le -f'(1 - \epsilon)\epsilon. \tag{90}$$

(ii) Let $\epsilon_0 \in (0,1)$ be fixed. For any $\epsilon \in [0,\epsilon_0]$ and any function $h : \mathbb{R}^p \to \mathbb{R}$, we have

$$K_f(P_\epsilon, P_{\theta^*}; h) \le -f'(1-\epsilon_0)\epsilon.$$

(iii) Suppose, in addition, that $f'(\mathrm{e}^u)$ is concave in u and $f^\#(\mathrm{e}^u)$ is R_1-Lipschitz in u as stated in Assumption 2. Let $\epsilon_1 \in (0,1)$ be fixed. If $\hat{\epsilon} = n^{-1}\sum_{i=1}^n U_i \in [0,\epsilon_1]$, then for any function $h : \mathbb{R}^p \to \mathbb{R}$,

$$K_f(P_n, P_{\theta^*}; h) \le -f'(1-\epsilon_1)\hat{\epsilon} + R_1|\mathrm{E}_{P_{\theta^*,n}} h(x) - \mathrm{E}_{P_{\theta^*}} h(x)|, \tag{91}$$

where $P_{\theta^*,n}$ denotes the empirical distribution of $\{X_i : U_i = 0, i = 1,\dots,n\}$ in the latent representation of Huber's contamination model.

Proof (i) Notice that by definition,

$$\begin{aligned}&(1-\epsilon)f'(t) - f^\#(t) = (1-\epsilon)f'(t) - tf'(t) + f(t)\\ &= f(t) + f'(t)((1-\epsilon)-t).\end{aligned}$$

By the convexity of f, we have that for any $t > 0$ and $\epsilon \in [0,1)$,

$$f(t) + f'(t)((1-\epsilon)-t) \le f(1-\epsilon).$$

Moreover, by the convexity of f and $f(1) = 0$, we have

$$f(1-\epsilon) \le f(1) + f'(1-\epsilon)((1-\epsilon)-1) = -f'(1-\epsilon)\epsilon.$$

Combining the preceding three displays yields the desired result.

(ii) For any function h, we have

$$\begin{aligned}&K_f(P_\epsilon, P_{\theta^*}; h) = \epsilon\,\mathrm{E}_Q f'(\mathrm{e}^{h(x)}) + \mathrm{E}_{P_{\theta^*}}\left\{(1-\epsilon)f'(\mathrm{e}^{h(x)}) - f^\#(\mathrm{e}^{h(x)})\right\}\\ &\le \mathrm{E}_{P_{\theta^*}}\left\{(1-\epsilon)f'(\mathrm{e}^{h(x)}) - f^\#(\mathrm{e}^{h(x)})\right\},\end{aligned} \tag{92}$$

using the fact that f is non-increasing and hence $f'(t) \le 0$ for $t > 0$. Setting $t = \mathrm{e}^{h(x)}$ in (90) shows that for $\epsilon \le \epsilon_0$,

$$(1-\epsilon)f'(\mathrm{e}^{h(x)}) - f^\#(\mathrm{e}^{h(x)}) \le -f'(1-\epsilon)\epsilon \le -f'(1-\epsilon_0)\epsilon. \tag{93}$$

where $f'(1-\epsilon) \ge f'(1-\epsilon_0)$ for $\epsilon \le \epsilon_0$ by the convexity of f. Combining (92) and (93) leads to the desired result.

(iii) For any function h, $K_f(P_n, P_{\theta^*,n}; h)$ can be bounded as follows:

$$\begin{aligned}&K_f(P_n, P_{\theta^*}; h)\\ &= \frac{1}{n}\sum_{i=1}^n U_i f'(\mathrm{e}^{h(X_i)}) + \frac{1}{n}\sum_{i=1}^n (1-U_i) f'(\mathrm{e}^{h(X_i)}) - \mathrm{E}_{P_{\theta^*}} f^\#(\mathrm{e}^{h(x)})\\ &\le (1-\hat{\epsilon})\mathrm{E}_{P_{\theta^*,n}} f'(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_{\theta^*}} f^\#(\mathrm{e}^{h(x)}) &(94)\\ &\le (1-\hat{\epsilon}) f'(\mathrm{e}^{\mathrm{E}_{P_{\theta^*,n}} h(x)}) - f^\#(\mathrm{e}^{\mathrm{E}_{P_{\theta^*}} h(x)}) &(95)\\ &\le -f'(1-\epsilon_1)\hat{\epsilon} + |f^\#(\mathrm{e}^{\mathrm{E}_{P_{\theta^*,n}} h(x)}) - f^\#(\mathrm{e}^{\mathrm{E}_{P_{\theta^*}} h(x)})| &(96)\\ &\le -f'(1-\epsilon_1)\hat{\epsilon} + R_1|\mathrm{E}_{P_{\theta^*,n}} h(x) - \mathrm{E}_{P_{\theta^*}} h(x)|. &(97)\end{aligned}$$

Line (94) follows because $f'(t) \le 0$ for $t > 0$. Line (95) follows from Jensen's inequality by the concavity of $f'(\mathrm{e}^u)$ and $-f^{\#}(\mathrm{e}^u)$ in u. Line (96) follows because

$$(1-\hat{\epsilon})f'(\mathrm{e}^{\mathrm{E}_{P_{\theta^*,n}} h(x)}) - f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\theta^*,n}} h(x)}) \le -f'(1-\epsilon_1)\hat{\epsilon},$$

obtained by taking $\epsilon = \hat{\epsilon}$ and $t = \mathrm{e}^{\mathrm{E}_{P_{\theta^*,n}} h(x)}$ in (90) and using $f'(1-\hat{\epsilon}) \ge f'(1-\epsilon_1)$ for $\hat{\epsilon} \le \epsilon_1$. Finally, line (97) follows because $f^{\#}(\mathrm{e}^u)$ is R_1-Lipschitz in u. ■

Remark 32 *Compared with (91), $K_f(P_n, P_{\theta^*}; h)$ can also be bounded as*

$$K_f(P_n, P_{\theta^*}; h) \le -f'(1-\epsilon_1)\hat{\epsilon} + |\mathrm{E}_{P_{\theta^*,n}} f^{\#}(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_{\theta^*}} f^{\#}(\mathrm{e}^{h(x)})|. \tag{98}$$

In fact, this follows directly from (94), because for $\hat{\epsilon} \le \epsilon_1$,

$$(1-\hat{\epsilon})f'(\mathrm{e}^{h(x)}) - f^{\#}(\mathrm{e}^{h(x)}) \le -f'(1-\epsilon_1)\hat{\epsilon},$$

which can be obtained by taking $\epsilon = \hat{\epsilon}$ and $t = \mathrm{e}^{h(x)}$ in (90), similarly as (93). However, the bound (98) involves the moment difference of $f^{\#}(\mathrm{e}^{h(x)})$ between P_{θ^*} and $P_{\theta^*,n}$, which is difficult to control for h in our spline class, even with $f^{\#}(\mathrm{e}^u)$ Lipschitz in u. In contrast, by exploiting the concavity of $f'(\mathrm{e}^u)$ and $-f^{\#}(\mathrm{e}^u)$ in u, the bound (91) is derived such that it involves the moment difference of $h(x)$, which can be controlled by Lemma 30 in Proposition 33 or by Lemma 43 in Proposition 44.

Proposition 33 *In the setting of Proposition 23, it holds with probability at least $1 - 5\delta$ that for any $\gamma \in \Gamma$,*

$$K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \le -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) + \mathrm{pen}_1(\gamma) C_{\mathrm{sp13}} R_1 M_{11} \lambda_{11},$$

where $C_{\mathrm{sp13}} = (5/3)(C_{\mathrm{sp11}} \vee C_{\mathrm{sp12}})$ with C_{sp11} and C_{sp12} as in Lemma 30, $M_{11} = M_1 + 2M_1^{1/2}\sqrt{2\pi}$, and

$$\lambda_{11} = \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{n}} + \frac{2\log(5p) + \log(\delta^{-1})}{n}.$$

Proof Consider the event $\Omega_1 = \{|\hat{\epsilon} - \epsilon| \le \sqrt{\epsilon(1-\epsilon)/(n\delta)}\}$. By Chebyshev's inequality, we have $\mathrm{P}(\Omega_1) \ge 1 - \delta$. In the event Ω_1, we have $|\hat{\epsilon} - \epsilon| \le 1/5$ by the assumption $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \le 1/5$ and hence $\hat{\epsilon} \le 2/5$ by the assumption $\epsilon \le 1/5$. By Lemma 31 with $\epsilon_1 = 2/5$, it holds in the event Ω_1 that for any $\gamma \in \Gamma$,

$$\begin{aligned}
&K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \\
&\le -f'(3/5)\hat{\epsilon} + R_1 \left| \mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\mu^*}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\mu^*}(x) \right| \\
&\le -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) + R_1 \left| \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x) \right|.
\end{aligned} \tag{99}$$

The last step (99) uses the fact that $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\mu^*}(x) = \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x)$ and $\mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\mu^*}(x) = \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*)$, by the definition $h_{\gamma,\mu^*}(x) = h_\gamma(x - \mu^*)$.

Next, conditionally on the contamination indicators $(U_1, \ldots, U_n)$ such that the event Ω_1 holds, we have that $\{X_i : U_i = 1, i = 1, \ldots, n\}$ are n_1 independent and identically distributed observations from P_{θ^*}, where $n_1 = \sum_{i=1}^n (1 - U_i) = n(1 - \hat{\epsilon}) \geq (3/5)n$. Denote as Ω_2 the event that for any γ_1 and γ_2,

$$\left| \mathrm{E}_{P_{\theta^*,n}} \gamma_1^{\mathrm{T}} \varphi(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} \gamma_1^{\mathrm{T}} \varphi(x) \right| \leq \|\gamma_1\|_1 C_{\mathrm{sp11}} M_1^{1/2} \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{(3/5)n}},$$

and

$$\begin{aligned} &\left| \mathrm{E}_{P_{\theta^*,n}} \gamma_2^{\mathrm{T}} (\varphi(x - \mu^*) \otimes \varphi(x - \mu^*)) - \mathrm{E}_{P_{(0,\Sigma^*)}} \gamma_2^{\mathrm{T}} (\varphi(x) \otimes \varphi(x)) \right| \\ &\leq \|\gamma_2\|_1 C_{\mathrm{sp12}} M_{11} \left\{ \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{(3/5)n}} + \frac{2\log(5p) + \log(\delta^{-1})}{(3/5)n} \right\}, \end{aligned}$$

where C_{sp11}, C_{sp12}, and M_{11} are defined as in Lemma 30 with $\|\xi\|_\infty = 2$. In the event Ω_2, the preceding inequalities imply that for any $\gamma = (\gamma_0, \gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}})^{\mathrm{T}} \in \Gamma$,

$$\begin{aligned} &\left| \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x) \right| \\ &\leq \mathrm{pen}_1(\gamma)(5/3)(C_{\mathrm{sp11}} \vee C_{\mathrm{sp12}}) M_{11} \lambda_{11}, \end{aligned} \tag{100}$$

where $h_\gamma(x) = \gamma_0 + \gamma_1^{\mathrm{T}} \varphi(x) + \gamma_2^{\mathrm{T}} (\varphi(x) \otimes \varphi(x))$ and $\mathrm{pen}_1(\gamma) = \|\gamma_1\|_1 + \|\gamma_2\|_1$. By applying Lemma 30 with $k = 5$ to $\{X_i - \mu^* : U_i = 1, i = 1, \ldots, n\}$, we have $\mathrm{P}(\Omega_2 | U_1, \ldots, U_n) \geq 1 - 4\delta$ for any $(U_1, \ldots, U_n)$ such that Ω_1 holds. Taking the expectation over $(U_1, \ldots, U_n)$ given Ω_1 shows that $\mathrm{P}(\Omega_2 | \Omega_1) \geq 1 - 4\delta$ and hence $\mathrm{P}(\Omega_1 \cap \Omega_2) \geq (1 - \delta)(1 - 4\delta) \geq 1 - 5\delta$.

Combining (99) and (100) in the event $\Omega_1 \cap \Omega_2$ indicates that, with probability at least $1 - 5\delta$, the desired inequality holds for any $\gamma \in \Gamma$. ■

Lemma 34 *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed as $X \sim P_\epsilon$. Let $b > 0$ be fixed and $g : \mathbb{R}^p \to [0,1]^q$ be a vector of fixed functions. For a convex and twice differentiable function $f : (0, \infty) \to \mathbb{R}$, define*

$$\begin{aligned} F(X_1, \ldots, X_n) &= \sup_{\|w\|_1 = 1, \mu \in \mathbb{R}^p} \left\{ K_f(P_n, P_{\hat{\theta}}; bw^{\mathrm{T}} g_\mu) - K_f(P_\epsilon, P_{\hat{\theta}}; bw^{\mathrm{T}} g_\mu) \right\} \\ &= \sup_{\|w\|_1 = 1, \mu \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n f'(\mathrm{e}^{bw^{\mathrm{T}} g_\mu(X_i)}) - \mathrm{E} f'(\mathrm{e}^{bw^{\mathrm{T}} g_\mu(X)}) \right\}, \end{aligned}$$

where $g_\mu(x) = g(x - \mu)$. Suppose that conditionally on $(X_1, \ldots, X_n)$, the random variable $Z_{n,j} = \sup_{\mu \in \mathbb{R}^p} |n^{-1} \sum_{i=1}^n \epsilon_i g_{\mu,j}(X_i)|$ is sub-gaussian with tail parameter $\sqrt{V_g/n}$ for $j = 1, \ldots, q$, where $(\epsilon_1, \ldots, \epsilon_n)$ are Rademacher variables, independent of $(X_1, \ldots, X_n)$, and $g_{\mu,j} : \mathbb{R}^p \to [0,1]$ denotes the jth component of g_μ. Then for any $\delta > 0$, we have that with probability at least $1 - 2\delta$,

$$F(X_1, \ldots, X_n) \leq b R_{2,b} \left\{ C_{\mathrm{sg6}} \sqrt{\frac{V_g \log(2q)}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}} \right\},$$

where $R_{2,b} = \sup_{|u| \leq b} \frac{\mathrm{d}}{\mathrm{d}u} f'(\mathrm{e}^u)$ and C_{sg6} is the universal constant in Lemma 70.

Proof First, F satisfies the bounded difference condition, because $|bw^{\mathrm{T}} g_\mu| \le b$ with $\|w\|_1 = 1$ and f' is non-decreasing by the convexity of f:

$$\begin{aligned}&\sup_{X_1,\ldots,X_n,X_i'} \left|F(X_1,\ldots,X_n) - F(X_1,\ldots,X_i',\ldots,X_n)\right| \\ &\le \frac{f'(\mathrm{e}^b) - f'(\mathrm{e}^{-b})}{n} \le \frac{2bR_{2,b}}{n},\end{aligned}$$

where $R_{2,b} = \sup_{|u|\le b} \frac{\mathrm{d}}{\mathrm{d}u} f'(\mathrm{e}^u)$. By McDiarmid's inequality (McDiarmid, 1989), for any $t > 0$, we have that with probability at least $1 - 2\mathrm{e}^{-2nt^2}$,

$$|F(X_1,\ldots,X_n) - \mathrm{E}F(X_1,\ldots,X_n)| \le 2bR_{2,b}t.$$

For any $\delta > 0$, taking $t = \sqrt{\log(\delta^{-1})/(2n)}$ shows that with probability at least $1 - 2\delta$,

$$|F(X_1,\ldots,X_n) - \mathrm{E}F(X_1,\ldots,X_n)| \le bR_{2,b}\sqrt{\frac{2\log(\delta^{-1})}{n}}.$$

Next, the expectation of $F(X_1,\ldots,X_n)$ can be bounded as follows:

$$\begin{aligned}&\mathrm{E} \sup_{\|w\|_1=1,\mu\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^n f'(\mathrm{e}^{bw^{\mathrm{T}} g_\mu(X_i)}) - \mathrm{E}f'(\mathrm{e}^{bw^{\mathrm{T}} g_\mu(X)}) \right\} \\ &\le 2\mathrm{E} \sup_{\|w\|_1=1,\mu\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^n \epsilon_i f'(\mathrm{e}^{bw^{\mathrm{T}} g_\mu(X_i)}) \right\} \qquad (101) \\ &\le 2R_{2,b}\,\mathrm{E} \sup_{\|w\|_1=1,\mu\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^n \epsilon_i bw^{\mathrm{T}} g_\mu(X_i) \right\} \qquad (102) \\ &\le 2bR_{2,b}\,\mathrm{E} \sup_{\mu\in\mathbb{R}^p} \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i g_\mu(X_i) \right\|_\infty \\ &\le 2bR_{2,b}C_{\mathrm{sg6}}\sqrt{\frac{V_g\log(2q)}{n}}. \qquad (103)\end{aligned}$$

Line (101) follows from the symmetrization Lemma 77, where $(\epsilon_1,\ldots,\epsilon_n)$ are Rademacher variables, independent of $(X_1,\ldots,X_n)$. Line (102) follows by Lemma 78, because $f'(\mathrm{e}^u)$ is $R_{2,b}$-Lipschitz in $u \in [-b,b]$. Line (103) follows because

$$\begin{aligned}&\mathrm{E} \sup_{\mu\in\mathbb{R}^p} \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i g_\mu(X_i) \right\|_\infty = \mathrm{E} \sup_{\mu\in\mathbb{R}^p} \max_{j=1,\ldots,q} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\mu,j}(X_i) \right| \\ &= \mathrm{E} \max_{j=1,\ldots,q} \sup_{\mu\in\mathbb{R}^p} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\mu,j}(X_i) \right| \\ &\le C_{\mathrm{sg6}}\sqrt{\frac{V_g\log(2q)}{n}}.\end{aligned}$$

For the last step, we use the assumption that conditionally on $(X_1,\ldots,X_n)$, the random variable $Z_{n,j} = \sup_{\mu\in\mathbb{R}^p} |n^{-1}\sum_{i=1}^n \epsilon_i g_{\mu,j}(X_i)|$ is sub-gaussian with tail parameter $\sqrt{V_g/n}$

for each $j = 1, \ldots, q$, and apply Lemma 70 to obtain $\mathrm{E}(\max_{j=1,\ldots,q} Z_{n,j} | X_1, \ldots, X_n) \leq C_{\mathrm{sg6}} \sqrt{V_g \log(2q)/n}$, and then $\mathrm{E}(\max_{j=1,\ldots,q} Z_{n,j}) \leq C_{\mathrm{sg6}} \sqrt{V_g \log(2q)/n}$.

Combining the tail probability and expectation bounds yields the desired result. ■

Remark 35 *Results of Lemma 34 and Lemma 45 still hold with $R_{2,b}$ and $R_{2,b\sqrt{q}}$ replaced by 1 if K_f is replaced by K_{HG}. This is true because $f'(\mathrm{e}^u)$ will be replaced by identity function which is just 1-Lipschitz.*

Lemma 36 *Suppose that $f : (0, \infty) \to \mathbb{R}$ is convex and three-times differentiable. Let $b > 0$ be fixed. For any function $h : \mathbb{R}^p \to [-b, b]$, we have*

$$K_f(P_\epsilon, P_{\hat{\theta}}; h) \geq f'(\mathrm{e}^{-b})\epsilon + f''(1) \left\{ \mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\} - \frac{1}{2} b^2 R_{3,b},$$

where $R_{3,b} = R_{31,b} + R_{32,b}$, $R_{31,b} = \sup_{|u| \leq b} \frac{\mathrm{d}^2}{\mathrm{d}u^2} \{-f'(\mathrm{e}^u)\}$, and $R_{32,b} = \sup_{|u| \leq b} \frac{\mathrm{d}^2}{\mathrm{d}u^2} f^{\#}(\mathrm{e}^u)$.

Proof First, $K_f(P_\epsilon, P_{\hat{\theta}}; h)$ can be bounded as

$$\begin{aligned}
&K_f(P_\epsilon, P_{\hat{\theta}}; h) \\
&= \epsilon \mathrm{E}_Q f'(\mathrm{e}^{h(x)}) + (1 - \epsilon) \mathrm{E}_{P_{\theta^*}} f'(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_{\hat{\theta}}} f^{\#}(\mathrm{e}^{h(x)}) \\
&\geq f'(\mathrm{e}^{-b})\epsilon + \mathrm{E}_{P_{\theta^*}} f'(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_{\hat{\theta}}} f^{\#}(\mathrm{e}^{h(x)}) \\
&= f'(\mathrm{e}^{-b})\epsilon + K_f(P_{\theta^*}, P_{\hat{\theta}}; h),
\end{aligned}$$

where the inequality follows because $f'(\mathrm{e}^{h(x)}) \geq f'(\mathrm{e}^{-b})$ for $h(x) \in [-b, b]$ by the convexity of f. Next, consider the function $\kappa(t) = K_f(P_{\theta^*}, P_{\hat{\theta}}; th)$. A Taylor expansion of $\kappa(1) = K_f(P_{\theta^*}, P_{\hat{\theta}}; h)$ about $t = 0$ yields

$$K_f(P_{\theta^*}, P_{\hat{\theta}}; h) = f''(1) \left\{ \mathrm{E}_{P_\theta} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\} - \frac{1}{2} \kappa''(t),$$

where for some $t \in [0, 1]$,

$$\kappa''(t) = -\mathrm{E}_{P_{\theta^*}} \left\{ h^2(x) \frac{\mathrm{d}^2}{\mathrm{d}u^2} f'(\mathrm{e}^u)|_{u=th(x)} \right\} + \mathrm{E}_{P_{\hat{\theta}}} \left\{ h^2(x) \frac{\mathrm{d}^2}{\mathrm{d}u^2} f^{\#}(\mathrm{e}^u)|_{u=th(x)} \right\}.$$

The desired result then follows because $h(x) \in [-b, b]$ and $th(x) \in [-b, b]$ for $t \in [0, 1]$, and hence $\kappa''(t) \leq R_{3,b}$ by the definition of $R_{3,b}$. ■

Proposition 37 *Let $b_1 > 0$ be fixed. In the setting of Proposition 23, it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_1(\gamma) = b_1$,*

$$\begin{aligned}
&K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\
&\geq f''(1) \left\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} + f'(\mathrm{e}^{-b_1})\epsilon - \frac{1}{2} b_1^2 R_{3,b_1} - b_1 R_{2,b_1} \lambda_{12}
\end{aligned}$$

where, with $C_{\mathrm{rad4}} = C_{\mathrm{sg6}} C_{\mathrm{rad3}}$,

$$\lambda_{12} = C_{\mathrm{rad4}} \sqrt{\frac{4\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}},$$

depending on the universal constants C_{sg6} *and* C_{rad3} *in Lemma 70 and Corollary 82.*

Proof By definition, for any $\gamma \in \Gamma_{\mathrm{rp}}$, $h_\gamma(x)$ can be represented as $h_{\mathrm{rp},\beta,c}(x)$ such that $\beta_0 = \gamma_0$ and $\mathrm{pen}_1(\beta) = \mathrm{pen}_1(\gamma)$:

$$h_{\mathrm{rp},\beta,c}(x) = \beta_0 + \sum_{j=1}^{p} \beta_{1j} \,\mathrm{ramp}(x_j - c_j) + \sum_{1 \le i \ne j \le p} \beta_{2,ij} \,\mathrm{ramp}(x_i)\mathrm{ramp}(x_j),$$

where $c = (c_1, \ldots, c_p)^{\mathrm{T}}$ with $c_j \in \{0, 1\}$, and $\beta = (\beta_0, \beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}}$ with $\beta_1 = (\beta_{1j} : j = 1, \ldots, p)^{\mathrm{T}}$ and $\beta_2 = (\beta_{2,ij} : 1 \le i \ne j \le p)$. Then for any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_1(\gamma) = b_1$, we have $\beta_0 = 0$ and $\mathrm{pen}_1(\beta) = b_1$ correspondingly, and hence $h_\gamma(x) = h_{\mathrm{rp},\beta,c}(x) \in [-b_1, b_1]$ by the boundedness of the ramp function in $[0, 1]$. Moreover, $h_{\mathrm{rp},\beta,c}(x)$ with $\beta_0 = 0$ and $\mathrm{pen}_1(\beta) = b_1$ can be expressed in the form $b_1 w^{\mathrm{T}} g(x)$, where for $q = 2p + p(p-1)$, $w \in \mathbb{R}^q$ is an L_1 unit vector, and $g : \mathbb{R}^p \to [0, 1]^q$ is a vector of functions including $\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j - 1)$ for $j = 1, \ldots, p$, and $\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$ for $1 \le i \ne j \le p$. For symmetry, $\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j)\mathrm{ramp}(x_i)$ are included as two distinct components in g, and the corresponding coefficients are identical to each other in w. Parenthetically, at most one of the coefficients in w associated with $\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j - 1)$ is nonzero for each j, but this property is not used in the subsequent discussion.

Next, $K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ can be bounded as

$$\begin{aligned} &K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ &\ge K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \{K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})\}. \end{aligned}$$

For any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_1(\gamma) = b_1$, applying Lemma 36 with $h = h_{\gamma,\hat{\mu}}$ and $b = b_1$ yields

$$\begin{aligned} &K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ &\ge f''(1) \left\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} + f'(\mathrm{e}^{-b_1})\epsilon - \frac{1}{2} b_1^2 R_{3,b_1}. \end{aligned}$$

By Lemma 34 with $b = b_1$ and $g(x) \in [0, 1]^q$ defined above, it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_1(\gamma) = b_1$,

$$\begin{aligned} &\{K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})\} \\ &\le b_1 R_{2,b_1} \left\{ C_{\mathrm{sg6}} \sqrt{\frac{V_g \log(2q)}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}} \right\} \\ &= b_1 R_{2,b_1} \left\{ C_{\mathrm{sg6}} C_{\mathrm{rad3}} \sqrt{\frac{4\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}} \right\}, \end{aligned}$$

where $V_g = 4C_{\mathrm{rad3}}^2$ is determined in Lemma 34 as follows. For $j = 1, \ldots, q$, consider the function class $\mathcal{G}_j = \{g_{\mu,j} : \mu \in \mathbb{R}^p\}$, where $\mu = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ and, as defined in Lemma 34,

$g_{\mu,j}(x)$ is either a moving-knot ramp function, $\mathrm{ramp}(x_{j_1} - \mu_{j_1})$ or $\mathrm{ramp}(x_{j_1} - \mu_{j_1} - 1)$ or a product of moving-knot ramp functions, $\mathrm{ramp}(x_{j_1} - \mu_{j_1})\mathrm{ramp}(x_{j_2} - \mu_{j_2})$ for $1 \le j_1 \ne j_2 \le p$. By Lemma 60, the VC index of moving-knot ramp functions is 2. By applying Corollary 82 (i) and (ii) with vanishing $\mathcal{H}$, we obtain that conditionally on $(X_1, \ldots, X_n)$, the random variable $Z_{n,j} = \sup_{\mu \in \mathbb{R}^p} |n^{-1} \sum_{i=1}^n \epsilon_i g_{\mu,j}(X_i)| = \sup_{f_j \in \mathcal{G}_j} |n^{-1} \sum_{i=1}^n \epsilon_i f_j(X_i)|$ is sub-gaussian with tail parameter $C_{\mathrm{rad}3}\sqrt{4/n}$ for $j = 1, \ldots, q$.

Combining the preceding three displays leads to the desired result. ■

Lemma 38 (Local linearity 1) *For $\delta \in \mathbb{R}$ and $0 \le \sigma_1, \sigma_2 \le M^{1/2}$, denote $D_h = \mathrm{E}h(\sigma_1 Z + \delta) - \mathrm{E}h(\sigma_2 Z)$, where h is a function on $\mathbb{R}$ and Z is a standard Gaussian random variable. For $h_1(x) = \pm\mathrm{ramp}(x)$, if $|D_{h_1}| \le a$ for $a \in (0, 1/2)$, then we have*

$$|\delta| \le S_{4,a}|D_{h_1}|, \tag{104}$$

where $S_{4,a} = (1 + \sqrt{2M \log \frac{2}{1-2a}})/a$. For $h_2(x) = \pm\mathrm{ramp}(x - 1)$, we have

$$|\sigma_1 - \sigma_2| \le S_5(|D_{h_2}| + |\delta|/2), \tag{105}$$

where $S_5 = 2\sqrt{2\pi}(1 - \mathrm{e}^{-2/M})^{-1}$.

Remark 39 *Define a ramp function class*

$$\mathcal{R}_1 = \{\pm\mathrm{ramp}(x - c), x \in \mathbb{R} : c = 0, 1\}.$$

In the setting of Lemma 38, suppose that for fixed $a \in (0, 1/2)$,

$$D \stackrel{def}{=} \sup_{h \in \mathcal{R}_1} \{\mathrm{E}h(\sigma_1 Z + \delta) - \mathrm{E}h(\sigma_2 Z)\} \le a.$$

Then we have

$$|\delta| \le S_{4,a} D, \quad |\sigma_1 - \sigma_2| \le S_{6,a} D,$$

where $S_{6,a} = S_5(1 + S_{4,a}/2)$. This shows that the moment matching discrepancy D over $\mathcal{R}_1$ delivers upper bounds, up to scaling constants, on the mean and standard deviation differences, provided that D is sufficiently small, for example, $D \le 1/3$.

Proof

[Proof of (104)] First, assume that δ is nonnegative. The other direction will be discussed later. Take $h(x) = \mathrm{ramp}(x)$. Then $h(x) + h(-x) = 1$ for all $x \in \mathbb{R}$ and

$$\mathrm{E}h(\sigma_2 Z) = \mathrm{E}h(\sigma_1 Z) = \frac{1}{2}.$$

Define $g(t) = \mathrm{E}h(\sigma_1 Z + t) - \mathrm{E}h(\sigma_1 Z) = \mathrm{E}h(\sigma_1 Z + t) - \frac{1}{2}$ for $t \in \mathbb{R}$. Then $g(0) = 0$ and $g(\delta) = D_h \le a$. We notice the following properties for the function g.

(i) $g(t)$ is non-decreasing and concave for $t \geq 0$.

(ii) $g(t)/t$ is non-increasing for $t > 0$.

(iii) $g(t) \geq \frac{1}{2} - \exp\{-(t-1)^2/(2\sigma_1^2)\}$ for $t \geq 1$.

For property (i), the derivative of $g(t)$ is $g'(t) = \frac{1}{2}\mathrm{E}\{\mathbb{1}_{\sigma_1 Z + t \in [-1,1]}\} \geq 0$. Moreover, $g'(t)$ is non-increasing: for $0 \leq t_1 < t_2$,

$$\begin{aligned} g'(t_1) &= \frac{1}{2}\mathrm{P}(-1 - t_1 \leq \sigma_1 Z \leq 1 - t_1) \\ &= \frac{1}{2}\mathrm{P}(-1 - t_1 \leq \sigma_1 Z \leq 1 - t_2) + \frac{1}{2}\mathrm{P}(1 - t_2 \leq \sigma_1 Z \leq 1 - t_1), \\ g'(t_2) &= \frac{1}{2}\mathrm{P}(-1 - t_2 \leq \sigma_1 Z \leq 1 - t_2) \\ &= \frac{1}{2}\mathrm{P}(-1 - t_1 \leq \sigma_1 Z \leq 1 - t_2) + \frac{1}{2}\mathrm{P}(-1 - t_2 \leq \sigma_1 Z \leq -1 - t_1), \end{aligned}$$

and

$$\begin{aligned} &\mathrm{P}(1 - t_2 \leq \sigma_1 Z \leq 1 - t_1) = \mathrm{P}(t_1 - 1 \leq \sigma_1 Z \leq t_2 - 1) \\ &\leq \mathrm{P}(t_1 + 1 \leq \sigma_1 Z \leq t_2 + 1) = \mathrm{P}(-1 - t_2 \leq \sigma_1 Z \leq -1 - t_1). \end{aligned}$$

The last inequality holds because $(t-1)^2 \leq (t+1)^2$ for any $t \geq 0$ and hence $\int_{t_1}^{t_2} \exp\{-(t-1)^2/(2\sigma_1^2)\}\,\mathrm{d}t \geq \int_{t_1}^{t_2} \exp\{-(t+1)^2/(2\sigma_1^2)\}\,\mathrm{d}t$. To show (ii), we write $g(t) = g(t) - g(0) = t\int_0^1 g'(tz)\,\mathrm{d}z$. Then $g(t)/t = \int_0^1 g'(tz)\,\mathrm{d}z$ is non-increasing in t because g' is non-increasing. To show (iii), we notice that $h(x) \geq \mathbb{1}_{x>1}$ and hence for $t \geq 1$,

$$\begin{aligned} g(t) + \frac{1}{2} &= \mathrm{E}h(\sigma_1 Z + t) \\ &\geq 1 - P(\sigma_1 Z + t \leq 1) = 1 - \mathrm{P}(\sigma_1 Z \geq t - 1) \\ &\geq 1 - \exp\{-(t-1)^2/(2\sigma_1^2)\}. \end{aligned}$$

The last inequality follows by the Gaussian tail bound: $\mathrm{P}(Z \geq z) \leq \mathrm{e}^{-z^2/2}$ for $z > 0$.

By the preceding properties, we show that $\delta \leq S_{4,a} D_h$ and hence (104) holds. Without loss of generality, assume that $\delta \neq 0$. For $a \in (0, 1/2)$, let $t_a > 0$ be determined such that $g(t_a) = a$. Then $\delta \leq t_a$ and, by property (ii), $g(\delta)/\delta \geq a/t_a$. If $t_a \geq 1$, then, by property (iii), $a = g(t_a) \geq \frac{1}{2} - \exp\{-(t_a - 1)^2/(2\sigma_1^2)\}$, and hence

$$t_a \leq 1 + \sqrt{2}\sigma_1 \sqrt{\log \frac{2}{1-2a}}.$$

This inequality remains valid if $t_a < 1$. Therefore, $\delta \leq g(\delta) t_a/a \leq g(\delta) S_{4,a} = D_h S_{4,a}$, by the assumption $\sigma_1 \leq M^{1/2}$ and the definition of $S_{4,a}$.

When δ is negative, a similar argument taking $h(x) = -\mathrm{ramp}(x)$, which is the same as $\mathrm{ramp}(-x) - 1$, shows that $-\delta \leq S_{4,a} D_h$ and hence (104) holds.

[Proof of (105)] First, assume that $\sigma_1 - \sigma_2$ is nonnegative. Take $h(x) = \mathrm{ramp}(x-1)$. Notice that $h(x)$ is (1/2)-Lipschitz and hence $|h(x+\delta) - h(x)| \le |\delta|/2$. Then by the triangle inequality, we have

$$\mathrm{E}h(\sigma_1 Z) - \mathrm{E}h(\sigma_2 Z) \le D_h + |\delta|/2.$$

Define $g(t) = \mathrm{E}h(tZ) - \mathrm{E}h(\sigma_2 Z)$ for $t \ge 0$. Then $g(\sigma_2) = 0$ and $g(\sigma_1) \le D_h + |\delta|/2$. The derivative $g'(t) = \frac{1}{2}\mathrm{E}\{Z \mathbb{1}_{tZ \in [0,2]}\}$ can be calculated as

$$g'(t) = \frac{1}{2}\int_0^{2/t} \frac{1}{\sqrt{2\pi}} z \mathrm{e}^{-z^2/2}\,\mathrm{d}z = \frac{1}{2\sqrt{2\pi}}\left\{1 - \mathrm{e}^{-\frac{(2/t)^2}{2}}\right\}.$$

By the mean value theorem, $g(\sigma_1) = g(\sigma_1) - g(\sigma_2) = g'(t)(\sigma_1 - \sigma_2) \ge g'(M^{1/2})(\sigma_1 - \sigma_2)$, where $t \in [\sigma_2, \sigma_1]$ and hence $t \le M^{1/2}$ because $\sigma_1 \le M^{1/2}$. Therefore, we have

$$\sigma_1 - \sigma_2 \le g'(M^{1/2})^{-1} g(\sigma_1) \le S_5(D_h + |\delta|/2),$$

by the definition $S_5 = g'(M^{1/2})^{-1} = 2\sqrt{2\pi}(1 - \mathrm{e}^{-2/M})^{-1}$.

When $\sigma_1 - \sigma_2$ is negative, a similar argument taking $h(x) = -\mathrm{ramp}(x-1)$ shows that $\sigma_2 - \sigma_1 \le S_5(D_h + |\delta|/2)$ and hence (105) holds. ■

Lemma 40 (Local linearity 2) *For $\delta_1, \delta_2 \in \mathbb{R}$, $0 \le \sigma_1, \sigma_2, \tilde{\sigma}_1, \tilde{\sigma}_2 \le M^{1/2}$, and $\rho, \tilde{\rho} \in [-1, 1]$, denote $D_h = \mathrm{E}h(\tilde{X}) - \mathrm{E}h(X)$, where h is a function on $\mathbb{R}^2$, $X = (X_1, X_2)^{\mathrm{T}}$ is a Gaussian random vector in $\mathbb{R}^2$ with mean 0 and variance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$, and $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)^{\mathrm{T}}$ is a Gaussian random vector in $\mathbb{R}^2$ with mean $\delta = (\delta_1, \delta_2)^{\mathrm{T}}$ and variance matrix $\begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{\sigma}_1\tilde{\sigma}_2\tilde{\rho} \\ \tilde{\sigma}_1\tilde{\sigma}_2\tilde{\rho} & \tilde{\sigma}_2^2 \end{pmatrix}$. For $h(x) = \pm\mathrm{ramp}(x_1)\mathrm{ramp}(x_2)$, we have*

$$|\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 - \rho\sigma_1\sigma_2| \le M^{1/2}\|\tilde{\sigma} - \sigma\|_1 + S_7(|D_h| + \Delta/2),$$

where $S_7 = 4\{(\frac{1}{\sqrt{2\pi M}}\mathrm{e}^{-1/(8M)}) \vee (1 - 2\mathrm{e}^{-1/(8M)})\}^{-2}$, which behaves like to $4(1-2\mathrm{e}^{-1/(8M)})^{-2}$ as $M \to 0$ or $8\pi M \mathrm{e}^{1/(4M)}$ as $M \to \infty$, and $\Delta = \|\delta\|_1 + \|\tilde{\sigma} - \sigma\|_1$, with $\sigma = (\sigma_1, \sigma_2)^{\mathrm{T}}$ and $\tilde{\sigma} = (\tilde{\sigma}_1, \tilde{\sigma}_2)^{\mathrm{T}}$.

Proof First, we handle the effect of different means and standard deviations between $\tilde{X}$ and X in D_h. Denote $D_h^\dagger = \mathrm{E}h(\tilde{Y}) - \mathrm{E}h(X)$, where $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2)^{\mathrm{T}}$ is a Gaussian random vector with mean 0 and variance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\tilde{\rho} \\ \sigma_1\sigma_2\tilde{\rho} & \sigma_2^2 \end{pmatrix}$. Then we have

$$|D_h^\dagger| \le |D_h| + \Delta/2. \tag{106}$$

In fact, assume that $\tilde{Y} = (\sigma_1 Z_1, \sigma_2 Z_2)^{\mathrm{T}}$ and $\tilde{X} = \delta + (\tilde{\sigma}_1 Z_1, \tilde{\sigma}_2 Z_2)^{\mathrm{T}}$, where $(Z_1, Z_2)^{\mathrm{T}}$ is a Gaussian random vector with mean 0 and variance matrix $\begin{pmatrix} 1 & \tilde{\rho} \\ \tilde{\rho} & 1 \end{pmatrix}$. For $h(x) =$

$\pm\text{ramp}(x_1)\text{ramp}(x_2)$, we have

$$\begin{aligned}
&|\mathrm{E}h(\tilde{Y}) - \mathrm{E}h(X)| \\
&\le |\mathrm{E}h(\tilde{X}) - \mathrm{E}h(X)| + \sum_{j=1,2} \mathrm{E}\,|\text{ramp}(\tilde{Y}_j) - \text{ramp}(\tilde{X}_j)| \\
&\le |\mathrm{E}h(\tilde{X}) - \mathrm{E}h(X)| + \sum_{j=1,2} \{\delta_j^2 + (\tilde{\sigma}_j - \sigma_j)^2\}^{1/2}/2 \\
&\le |\mathrm{E}h(\tilde{X}) - \mathrm{E}h(X)| + \Delta/2.
\end{aligned}$$

The first inequality follows by the triangle inequality and the fact that $\text{ramp}(\cdot)$ is bounded in $[0,1]$. The second step uses $\mathrm{E}\,|\text{ramp}(\tilde{Y}_j) - \text{ramp}(\tilde{X}_j)| \le (1/2)\mathrm{E}|\tilde{Y}_j - \tilde{X}_j| \le (1/2)\mathrm{E}^{1/2}[\{\delta_j + (\tilde{\sigma}_j - \sigma_j)Z_j\}^2] = \{\delta_j^2 + (\tilde{\sigma}_j - \sigma_j)^2\}^{1/2}/2$, by the fact that $\text{ramp}(\cdot)$ is $(1/2)$-Lipschitz, $\mathrm{E}Z_j = 0$, and $\mathrm{E}(Z_j^2) = 1$. The third inequality follows because $\sqrt{u_1 + u_2} \le \sqrt{u_1} + \sqrt{u_2}$.

Next, we show that

$$|\tilde{\rho} - \rho|\sigma_1\sigma_2 \le 8\pi M \mathrm{e}^{1/(4M)}|D_h^\dagger|. \tag{107}$$

Assume that $\tilde{\rho} - \rho$ is nonnegative. The other direction will be discussed later. Now take $h(x) = \text{ramp}(x_1)\text{ramp}(x_2)$, and define $g(t) = \mathrm{E}h(Y) - \mathrm{E}h(X)$ for $t \in [-1,1]$, where $Y = (Y_1, Y_2)^{\mathrm{T}}$ is a Gaussian random vector with mean 0 and variance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 t \\ \sigma_1\sigma_2 t & \sigma_2^2 \end{pmatrix}$. Then $g(\rho) = 0$ and $g(\tilde{\rho}) = D_h^\dagger$. The derivative $g'(t)$ can be calculated as

$$g'(t) = \frac{1}{4}\sigma_1\sigma_2 \mathrm{E}\mathbb{1}_{Y_1 \in [-1,1]}\mathbb{1}_{Y_2 \in [-1,1]}. \tag{108}$$

In fact, Y_1 can be represented as $Y_1 = t\frac{\sigma_1}{\sigma_2}Y_2 + \sqrt{1-t^2}\sigma_1 Z_1$, where Z_1 is a standard Gaussian variable independent of Y_2. Then direct calculation yields

$$\begin{aligned}
g'(t) &= \frac{1}{2}\mathrm{E}\mathbb{1}_{Y_1 \in [-1,1]}\left(\frac{\sigma_1}{\sigma_2}Y_2 - \frac{t}{\sqrt{1-t^2}}\sigma_1 Z_1\right)\text{ramp}(Y_2) \\
&= \frac{1}{2}\mathrm{E}\mathbb{1}_{Y_1 \in [-1,1]}\left\{\frac{\sigma_1}{\sigma_2}Y_2 - \frac{t}{1-t^2}(Y_1 - t\frac{\sigma_1}{\sigma_2}Y_2)\right\}\text{ramp}(Y_2) \\
&= \frac{1}{2}\mathrm{E}\mathbb{1}_{Y_1 \in [-1,1]}\left(\frac{1}{1-t^2}\frac{\sigma_1}{\sigma_2}Y_2 - \frac{t}{1-t^2}Y_1\right)\text{ramp}(Y_2) \\
&= \frac{1}{2}\frac{1}{1-t^2}\frac{\sigma_1}{\sigma_2}\mathrm{E}\mathbb{1}_{Y_1 \in [-1,1]}\left(Y_2 - t\frac{\sigma_2}{\sigma_1}Y_1\right)\text{ramp}(Y_2).
\end{aligned}$$

By Stein's lemma using the fact that Y_2 given Y_1 is Gaussian with mean $t\frac{\sigma_2}{\sigma_1}Y_1$ and variance $(1-t^2)\sigma_2^2$, we have

$$\mathrm{E}\left\{\left(Y_2 - t\frac{\sigma_2}{\sigma_1}Y_1\right)\text{ramp}(Y_2)\Big|Y_1\right\} = \mathrm{E}\left\{(1-t^2)\sigma_2^2\frac{1}{2}\mathbb{1}_{Y_2 \in [-1,1]}\Big|Y_1\right\}.$$

Substituting this into the expression for $g'(t)$ gives the formula (108).

To show (107), we derive a lower bound on $g'(t)$. Assume without loss of generality that $\sigma_1 \le \sigma_2$, because formula (108) is symmetric in Y_1 and Y_2. By the previous representation of Y_1, we have $|Y_1| \le \sqrt{1-t^2}\sigma_1|Z_1| + \frac{\sigma_1}{\sigma_2}|Y_2| \le \sigma_1|Z_1| + |Y_2|$ and hence

$$\begin{aligned} g'(t) &\ge \frac{1}{4}\sigma_1\sigma_2 \mathrm{E}\mathbb{1}_{Y_1\in[-1,1]}\mathbb{1}_{Y_2\in[-1/2,1/2]} \\ &\ge \frac{1}{4}\sigma_1\sigma_2 \mathrm{E}\mathbb{1}_{\sigma_1|Z_1|\in[-1/2,1/2]}\mathbb{1}_{Y_2\in[-1/2,1/2]} \\ &\ge S_7^{-1}\sigma_1\sigma_2. \end{aligned}$$

where $S_7 = 4\{(\frac{1}{\sqrt{2\pi M^{1/2}}}\mathrm{e}^{-1/(8M)}) \vee (1 - 2\mathrm{e}^{-1/(8M)})\}^{-2}$. The last inequality follows by the independence of Z_1 and Y_2, $\sigma_1 \le M^{1/2}$, $\sigma_2 \le M^{1/2}$, and the two probability bounds: $\mathrm{P}(M^{1/2}|Z_1| \le 1/2) \ge \frac{1}{\sqrt{2\pi M}}\mathrm{e}^{-(1/2)^2/(2M)}$ and $\mathrm{P}(M^{1/2}|Z_1| \le 1/2) \ge 1 - 2\mathrm{e}^{-(1/2)^2/(2M)}$. Hence by the mean value theorem, we have

$$g(\tilde{\rho}) = g(\tilde{\rho}) - g(\rho) \ge (\tilde{\rho} - \rho)\sigma_1\sigma_2 S_7^{-1},$$

which gives the desired bound (107) in the case of $\tilde{\rho} \ge \rho$:

$$(\tilde{\rho} - \rho)\sigma_1\sigma_2 \le S_7 g(\tilde{\rho}) = S_7 D_h^{\dagger}.$$

When $\tilde{\rho} - \rho$ is negative, a similar argument taking $h(x) = -\mathrm{ramp}(x_1)\mathrm{ramp}(x_2)$ and interchanging the roles of $\tilde{Y}$ and X leads to (107).

Finally, by the triangle inequality, we find

$$\begin{aligned} |\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 - \rho\sigma_1\sigma_2| &\le |\tilde{\sigma}_1\tilde{\sigma}_2 - \sigma_1\sigma_2| + |\tilde{\rho} - \rho|\sigma_1\sigma_2 \\ &\le M^{1/2}(|\tilde{\sigma}_1 - \sigma_1| + |\tilde{\sigma}_2 - \sigma_2|) + |\tilde{\rho} - \rho|\sigma_1\sigma_2. \end{aligned}$$

Combining this with (106) and (107) leads to the desired result. ■

Remark 41 *Define a ramp main-effect and interaction class*

$$\begin{aligned} \mathcal{R}_2 = \big\{ &\pm\mathrm{ramp}(x_j - c), (x_1, x_2) \in \mathbb{R}^2 : j = 1, 2,\, c = 0, 1 \big\} \\ &\bigcup \big\{ \pm\mathrm{ramp}(x_1)\mathrm{ramp}(x_2), (x_1, x_2) \in \mathbb{R}^2 \big\}. \end{aligned}$$

In the setting of Lemma 40, suppose that for fixed $a \in (0, 1/2)$,

$$D \overset{def}{=} \sup_{h\in\mathcal{R}_2} \Big\{ \mathrm{E}h(\tilde{X}) - \mathrm{E}h(X) \Big\} \le a.$$

Then combining Lemma 38–40 yields

$$\begin{aligned} &\max(|\delta_1|, |\delta_2|) \le S_{4,a} D, \quad \max(|\tilde{\sigma}_1 - \sigma_1|, |\tilde{\sigma}_2 - \sigma_2|) \le S_{6,a} D, \\ &\max(|\tilde{\sigma}_1^2 - \sigma_1^2|, |\tilde{\sigma}_2^2 - \sigma_2^2|, |\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 - \rho\sigma_1\sigma_2|) \le S_{8,a} D \end{aligned}$$

where $S_{8,a} = S_7(1 + S_{4,a} + S_{6,a}) + 2M^{1/2}S_{6,a}$. This shows that the moment matching discrepancy D over $\mathcal{R}_2$ delivers upper bounds, up to scaling constants, on the mean, variance, and covariance differences, provided that D is sufficiently small.

Proposition 42 *In the setting of Proposition 23 or Proposition 21, suppose that for $a \in (0, 1/2)$,*

$$D \stackrel{def}{=} \sup_{\gamma \in \Gamma_{\mathrm{rp}}, \mathrm{pen}_1(\gamma)=1} \left\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} \leq a. \tag{109}$$

Then we have

$$\|\hat{\mu} - \mu^*\|_\infty \leq S_{4,a} D,$$
$$\|\hat{\Sigma} - \Sigma^*\|_{\max} \leq S_{8,a} D,$$

where $S_{4,a}$ and $S_{8,a}$ are defined as in Lemma 38 and Remark 41 with $M = M_1$.

Proof For any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\mathrm{pen}_1(\gamma) = 1$, the function $h_{\gamma,\hat{\mu}}(x) = h_\gamma(x - \hat{\mu})$ can be expressed as $h_{\mathrm{rp},\beta,c}(x - \hat{\mu})$ with $\mathrm{pen}_1(\beta) = 1$. For $j = 1, \ldots, p$, by restricting $h_{\gamma,\hat{\mu}}(x)$ in (109) to those with $h_\gamma(x)$ defined as ramp functions of x_j and using Remark 39, we obtain

$$|\hat{\mu}_j - \mu_j^*| \leq S_{4,a} D, \quad |\hat{\sigma}_j - \sigma_j^*| \leq S_{6,a} D.$$

For $1 \leq i \neq j \leq p$, by restricting $h_{\gamma,\hat{\mu}}(x)$ in (109) to those with $h_\gamma(x)$ defined as ramp interaction functions of (x_i, x_j) and using Remark 41, we obtain

$$\max\left(|\hat{\Sigma}_{ii} - \Sigma_{ii}^*|, |\hat{\Sigma}_{jj} - \Sigma_{jj}^*|, |\hat{\Sigma}_{ij} - \Sigma_{ij}^*| \right) \leq S_{8,a} D,$$

where $\hat{\Sigma}_{ij}$ and Σ_{ij}^* are the (i,j)th elements of $\hat{\Sigma}$ and Σ^* respectively. Combining the preceding two displays leads to the desired result. ■

C.3 Details in main proof of Theorem 12

Lemma 43 *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed as $X \sim N_p(0, \Sigma)$ with $\|\Sigma\|_{\mathrm{op}} \leq M_2$. For k fixed knots $\xi_1, \ldots, \xi_k$ in $\mathbb{R}$, denote $\varphi(x) = (\varphi_1^{\mathrm{T}}(x), \ldots, \varphi_k^{\mathrm{T}}(x))^{\mathrm{T}}$, where $\varphi_l(x) \in \mathbb{R}^p$ is obtained by applying $t \mapsto (t - \xi_l)_+$ componentwise to $x \in \mathbb{R}^p$ for $l = 1, \ldots, k$. Then the following results hold.*

(i) $\varphi(X) - \mathrm{E}\varphi(X)$ is a sub-gaussian random vector with tail parameter $(kM_2)^{1/2}$.

(ii) For any $\delta > 0$, we have that with probability at least $1 - \delta$,

$$\sup_{\|w\|_2=1} \left| \frac{1}{n} \sum_{i=1}^n w^{\mathrm{T}} \varphi(X_i) - \mathrm{E} w^{\mathrm{T}} \varphi(X) \right|$$
$$\leq C_{\mathrm{sp21}} (kM_2)^{1/2} \sqrt{\frac{kp + \log(\delta^{-1})}{n}},$$

where $C_{\mathrm{sp21}} = \sqrt{2} C_{\mathrm{sg7}} C_{\mathrm{sg5}}$ and $(C_{\mathrm{sg5}}, C_{\mathrm{sg7}})$ are the universal constants in Lemmas 69 and 71.

(iii) Let $\mathcal{A}_2 = \{A \in \mathbb{R}^{kp\times kp} : \|A\|_{\mathrm{F}} = 1, A^{\mathrm{T}} = A\}$. *For any* $\delta > 0$, *we have that with probability at least* $1 - 3\delta$, *both the inequality in (ii) and*

$$\frac{1}{\sqrt{kp}} \sup_{A\in\mathcal{A}_2} \left| \frac{1}{n}\sum_{i=1}^{n} \varphi^{\mathrm{T}}(X_i)A\varphi(X_i) - \mathrm{E}\varphi^{\mathrm{T}}(X)A\varphi(X) \right| \leq C_{\mathrm{sp}22} k M_{21} \left\{ \sqrt{\frac{kp + \log(\delta^{-1})}{n}} + \frac{kp + \log(\delta^{-1})}{n} \right\},$$

where $M_{21} = M_2^{1/2}(M_2^{1/2} + \sqrt{2\pi}\|\xi\|_\infty)$, $\|\xi\|_\infty = \max_{l=1,\ldots,k} |\xi_l|$, $C_{\mathrm{sp}22} = \sqrt{2/\pi}C_{\mathrm{sp}21} + C_{\mathrm{sg}8}$, *and* $C_{\mathrm{sg}8}$ *is the universal constant in Lemma 72.*

Proof (i) First, we show that $w^{\mathrm{T}}\varphi(x)$ is a $k^{1/2}$-Lipschitz function for any L_2 unit vector $w \in \mathbb{R}^{kp}$. For any $x_1, x_2 \in \mathbb{R}^p$, we have $|w^{\mathrm{T}}(\varphi(x_1) - \varphi(x_2))| \leq \sum_{l=1}^{k} \|w_l\|_2 \|\varphi_l(x_1) - \varphi_l(x_2)\|_2 \leq \sum_{l=1}^{k} \|w_l\|_2 \|x_1 - x_2\| \leq k^{1/2}\|x_1 - x_2\|$, where w is partitioned as $w = (w_1^{\mathrm{T}}, \ldots, w_k^{\mathrm{T}})^{\mathrm{T}}$, and $\|\varphi_l(x_1) - \varphi_l(x_2)\|_2 \leq \|x_1 - x_2\|_2$ because each component of $\varphi_l(x)$ is 1-Lipschitz, as a function of only the corresponding component of x. Next, X can be represented as $\Sigma^{1/2}Z$, where Z is a standard Gaussian random vector. For any $z_1, z_2 \in \mathbb{R}^p$ and L_2 unit vector $w \in \mathbb{R}^{kp}$, we have

$$\begin{aligned} &|w^{\mathrm{T}}\varphi(\Sigma^{1/2}z_1) - w^{\mathrm{T}}\varphi(\Sigma^{1/2}z_2)| \\ &\leq k^{1/2}\|\Sigma^{1/2}(z_1 - z_2)\|_2 \leq k^{1/2}\|\Sigma^{1/2}\|_{\mathrm{op}}\|z_1 - z_2\|_2 \leq (kM_2)^{1/2}\|z_1 - z_2\|_2. \end{aligned}$$

Hence $w^{\mathrm{T}}\varphi(X)$ is a $(kM_2)^{1/2}$-Lipschitz function of the standard Gaussian vector Z. By Theorem 5.6 in Boucheron et al. (2013), the centered version satisfies that for any $t > 0$,

$$\mathrm{P}\left(|w^{\mathrm{T}}(\varphi(X) - \mathrm{E}\varphi(X))| > t\right) \leq 2\mathrm{e}^{-t^2/(2kM_2)}.$$

That is, $w^{\mathrm{T}}(\varphi(X) - \mathrm{E}\varphi(X))$ is sub-gaussian with tail parameter $(kM_2)^{1/2}$. The desired result follows by the definition of sub-gaussian random vectors.

(ii) As shown above, $w^{\mathrm{T}}(\varphi(X) - \mathrm{E}\varphi(X))$ is sub-gaussian with tail parameter $(kM_2)^{1/2}$ for any L_2 unit vector w. Then $w^{\mathrm{T}}\{n^{-1}\sum_{i=1}^{n}\varphi(X_i) - \mathrm{E}\varphi(X)\}$ is sub-gaussian with tail parameter $C_{\mathrm{sg}5}(kM_2/n)^{1/2}$ by sub-gaussian concentration (Lemma 69). Hence by definition, we have that $n^{-1}\sum_{i=1}^{n}\varphi(X_i) - \mathrm{E}\varphi(X)$ is a sub-gaussian random vector with tail parameter $C_{\mathrm{sg}5}(kM_2/n)^{1/2}$. Notice that

$$\sup_{\|w\|_2=1} \left| \frac{1}{n}\sum_{i=1}^{n} w^{\mathrm{T}}\varphi(X_i) - \mathrm{E}w^{\mathrm{T}}\varphi(X) \right| = \left\| \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) - \mathrm{E}\varphi(X) \right\|_2.$$

The desired result follows from Lemma 71: with probability at least $1 - \delta$, we have

$$\begin{aligned} \left\| \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) - \mathrm{E}\varphi(X) \right\|_2 &\leq C_{\mathrm{sg}7}C_{\mathrm{sg}5}(kM_2)^{1/2} \left\{ \sqrt{\frac{kp}{n}} + \sqrt{\frac{\log(\delta^{-1})}{n}} \right\} \\ &\leq \sqrt{2}C_{\mathrm{sg}7}C_{\mathrm{sg}5}(kM_2)^{1/2}\sqrt{\frac{kp + \log(\delta^{-1})}{n}}. \end{aligned}$$

(iii) The difference of interest can be expressed in terms of the centered variables as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\varphi_i^{\mathrm{T}}A\varphi_i - \mathrm{E}\varphi^{\mathrm{T}}A\varphi \\
&= \frac{1}{n}\sum_{i=1}^{n}(\varphi_i - \mathrm{E}\varphi)^{\mathrm{T}}A(\varphi_i - \mathrm{E}\varphi) - \mathrm{E}\{(\varphi - \mathrm{E}\varphi)^{\mathrm{T}}A(\varphi - \mathrm{E}\varphi)\} \qquad (110)\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}2(\mathrm{E}\varphi)^{\mathrm{T}}A(\varphi_i - \mathrm{E}\varphi). \qquad (111)
\end{aligned}$$

We handle the concentration of the two terms separately. Denote $\varphi_i = \varphi(X_i)$, $\varphi = \varphi(X)$, $\tilde{\varphi}_i = \varphi_i - \mathrm{E}\varphi$, and $\tilde{\varphi} = \varphi - \mathrm{E}\varphi$.

First, for $A \in \mathcal{A}_2$ the term in (111) can be bounded as follows:

$$\begin{aligned}
&\left|2(\mathrm{E}\varphi)^{\mathrm{T}}A\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\right| \le 2\|\mathrm{E}\varphi\|_2\|A\|_{\mathrm{op}}\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\right\|_2 \le 2\|\mathrm{E}\varphi\|_2\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\right\|_2 \\
&\le 2\sqrt{kp}\left(\frac{M_2^{1/2}}{\sqrt{2\pi}} + \|\xi\|_\infty\right)\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\right\|_2,
\end{aligned}$$

where $\|\xi\|_\infty = \max_{l=1,\ldots,k}|\xi_l|$. The second inequality holds because $\|A\|_{\mathrm{op}} \le \|A\|_{\mathrm{F}} = 1$. The third inequality holds because $\|\mathrm{E}\varphi_l\|_2 \le \sqrt{p}(M_2^{1/2}/\sqrt{2\pi} + |\xi_l|)$ by Lemma 59. By (ii), for any $\delta > 0$, we have that with probability at least $1-\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\right\|_2 \le C_{\mathrm{sp21}}(kM_2)^{1/2}\sqrt{\frac{kp + \log(\delta^{-1})}{n}}.$$

From the preceding two displays, we obtain that with probability at least $1-\delta$,

$$\begin{aligned}
&\frac{1}{\sqrt{kp}}\sup_{A\in\mathcal{A}_2}\left|2(\mathrm{E}\varphi)^{\mathrm{T}}A\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\right| \\
&\le \sqrt{\frac{2}{\pi}}C_{\mathrm{sp21}}(kM_2)^{1/2}\left(M_2^{1/2} + \sqrt{2\pi}\|\xi\|_\infty\right)\sqrt{\frac{kp + \log(\delta^{-1})}{n}}. \qquad (112)
\end{aligned}$$

Next, consider an eigen-decomposition $A = \sum_{l=1}^{kp}\lambda_l w_l w_l^{\mathrm{T}}$, where λ_l's are eigenvalues and w_l's are the eigenvectors with $\|w_l\|_2 = 1$. The concentration of the term in (110) can be controlled as follows:

$$\begin{aligned}
&\sup_{A\in\mathcal{A}_2}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i^{\mathrm{T}}A\tilde{\varphi}_i - \mathrm{E}\tilde{\varphi}^{\mathrm{T}}A\tilde{\varphi}\right| = \sup_{A\in\mathcal{A}_2}\left|\sum_{l=1}^{kp}\lambda_l w_l^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\tilde{\varphi}_i^{\mathrm{T}} - \mathrm{E}\tilde{\varphi}\tilde{\varphi}^{\mathrm{T}}\right)w_l\right| \\
&\le \sup_{A\in\mathcal{A}_2}\left(\sum_{l=1}^{kp}|\lambda_l|\right)\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\tilde{\varphi}_i^{\mathrm{T}} - \mathrm{E}\tilde{\varphi}\tilde{\varphi}^{\mathrm{T}}\right\|_{\mathrm{op}} \le \sqrt{kp}\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}_i\tilde{\varphi}_i^{\mathrm{T}} - \mathrm{E}\tilde{\varphi}\tilde{\varphi}^{\mathrm{T}}\right\|_{\mathrm{op}}.
\end{aligned}$$

The last inequality uses the fact that $\|A\|_{\mathrm{F}} = (\sum_{l=1}^{kp} \lambda_l^2)^{1/2} = 1$ and hence $\sum_{l=1}^{kp} |\lambda_l| \le \sqrt{kp}$ for $A \in \mathcal{A}_2$. From (i), $\tilde{\varphi}_i$ is a sub-gaussian random vector with tail parameter $(kM_2)^{1/2}$. By Lemma 72, for any $\delta > 0$, we have that with probability at least $1 - 2\delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \tilde{\varphi}_i^{\mathrm{T}} - \mathrm{E} \tilde{\varphi} \tilde{\varphi}^{\mathrm{T}} \right\|_{\mathrm{op}} \le C_{\mathrm{sg8}} k M_2 \left\{ \sqrt{\frac{kp + \log(\delta^{-1})}{n}} + \frac{kp + \log(\delta^{-1})}{n} \right\},$$

From the preceding two displays, we obtain that with probability at least $1 - 2\delta$,

$$\begin{aligned} &\frac{1}{\sqrt{kp}} \sup_{A \in \mathcal{A}_2} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i^{\mathrm{T}} A \tilde{\varphi}_i - \mathrm{E} \tilde{\varphi}^{\mathrm{T}} A \tilde{\varphi} \right| \\ &\le C_{\mathrm{sg8}} k M_2 \left\{ \sqrt{\frac{kp + \log(\delta^{-1})}{n}} + \frac{kp + \log(\delta^{-1})}{n} \right\}. \end{aligned} \tag{113}$$

Combining the two bounds (112) and (113) gives the desired result. ■

Proposition 44 *In the setting of Proposition 24, it holds with probability at least* $1 - 4\delta$ *that for any* $\gamma = (\gamma_0, \gamma_1, \gamma_2)^{\mathrm{T}} \in \Gamma$,

$$\begin{aligned} K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \le &-f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) \\ &+ \mathrm{pen}_2(\gamma_1)(5/3) C_{\mathrm{sp21}} M_2^{1/2} R_1 \lambda_{21} \\ &+ \mathrm{pen}_2(\gamma_2)(25\sqrt{5}/3) C_{\mathrm{sp22}} M_{21} R_1 \sqrt{p} \lambda_{31}, \end{aligned}$$

where C_{sp21} *and* C_{sp22} *are defined as in Lemma 43,* $M_{21} = M_2^{1/2}(M_2^{1/2} + 2\sqrt{2\pi})$, *and*

$$\lambda_{21} = \sqrt{\frac{5p + \log(\delta^{-1})}{n}}, \quad \lambda_{31} = \lambda_{21} + \frac{5p + \log(\delta^{-1})}{n}.$$

Proof Consider the event $\Omega_1 = \{|\hat{\epsilon} - \epsilon| \le \sqrt{\epsilon(1-\epsilon)/(n\delta)}\}$. By Chebyshev's inequality, we have $\mathrm{P}(\Omega_1) \ge 1 - \delta$. In the event Ω_1, we have $|\hat{\epsilon} - \epsilon| \le 1/5$ by the assumption $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \le 1/5$ and hence $\hat{\epsilon} \le 2/5$ by the assumption $\epsilon \le 1/5$. By Lemma 31 with $\epsilon_1 = 2/5$, it holds in the event Ω_1 that for any $\gamma \in \Gamma$,

$$\begin{aligned} &K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \\ &\le -f'(3/5)\hat{\epsilon} + R_1 \left| \mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\mu^*}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\mu^*}(x) \right| \\ &\le -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) + R_1 \left| \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x) \right|. \end{aligned} \tag{114}$$

The last step also uses the fact that $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\mu^*}(x) = \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x)$ and also $\mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\mu^*}(x) = \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*)$, by the definition $h_{\gamma,\mu^*}(x) = h_\gamma(x - \mu^*)$.

Next, conditionally on the contamination indicators $(U_1, \ldots, U_n)$ such that the event Ω_1 holds, we have that $\{X_i : U_i = 1, i = 1, \ldots, n\}$ are n_1 independent and identically

distributed observations from P_{θ^*}, where $n_1 = \sum_{i=1}^n (1 - U_i) = n(1 - \hat{\epsilon}) \geq (3/5)n$. Denote as Ω_2 the event that for any γ_1 and γ_2,

$$\left| \mathrm{E}_{P_{\theta^*,n}} \gamma_1^{\mathrm{T}} \varphi(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} \gamma_1^{\mathrm{T}} \varphi(x) \right| \leq \|\gamma_1\|_2 C_{\mathrm{sp}21} M_2^{1/2} \sqrt{\frac{5p + \log(\delta^{-1})}{(3/5)n}},$$

and

$$\begin{aligned} &\left| \mathrm{E}_{P_{\theta^*,n}} \gamma_2^{\mathrm{T}} (\varphi(x - \mu^*) \otimes \varphi(x - \mu^*)) - \mathrm{E}_{P_{(0,\Sigma^*)}} \gamma_2^{\mathrm{T}} (\varphi(x) \otimes \varphi(x)) \right| \\ &\leq \|\gamma_2\|_2 C_{\mathrm{sp}22} 5 M_{21} \sqrt{5p} \left\{ \sqrt{\frac{5p + \log(\delta^{-1})}{(3/5)n}} + \frac{5p + \log(\delta^{-1})}{(3/5)n} \right\}, \end{aligned}$$

where $C_{\mathrm{sp}21}$, $C_{\mathrm{sp}22}$, and M_{21} are defined as in Lemma 43 with $\|\xi\|_\infty = 1$. In the event Ω_2, the preceding inequalities imply that for any $\gamma = (\gamma_0, \gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}})^{\mathrm{T}} \in \Gamma$,

$$\begin{aligned} &\left| \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x) \right| \\ &\leq \mathrm{pen}_2(\gamma_1)(5/3) C_{\mathrm{sp}21} M_2^{1/2} \lambda_{21} + \mathrm{pen}_2(\gamma_2)(5/3) C_{\mathrm{sp}22} 5 M_{21} \sqrt{5p} \lambda_{31}, \end{aligned} \tag{115}$$

where $h_\gamma(x) = \gamma_0 + \gamma_1^{\mathrm{T}} \varphi(x) + \gamma_2^{\mathrm{T}} (\varphi(x) \otimes \varphi(x))$, $\mathrm{pen}_2(\gamma_1) = \|\gamma_1\|_2$, and $\mathrm{pen}_2(\gamma_2) = \|\gamma_2\|_2$. By applying Lemma 43 with $k = 5$ to $\{X_i - \mu^* : U_i = 1, i = 1, \ldots, n\}$, we have $\mathrm{P}(\Omega_2 | U_1, \ldots, U_n) \geq 1 - 3\delta$ for any $(U_1, \ldots, U_n)$ such that Ω_1 holds. Taking the expectation over $(U_1, \ldots, U_n)$ given Ω_1 shows that $\mathrm{P}(\Omega_2 | \Omega_1) \geq 1 - 3\delta$ and hence $\mathrm{P}(\Omega_1 \cap \Omega_2) \geq (1 - \delta)(1 - 3\delta) \geq 1 - 4\delta$.

Combining (114) and (115) in the event $\Omega_1 \cap \Omega_2$ indicates that, with probability at least $1 - 4\delta$, the desired inequality holds for any $\gamma \in \Gamma$. ■

Lemma 45 *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed as $X \sim P_\epsilon$. Let $b > 0$ be fixed and $g : \mathbb{R}^p \to [0,1]^q$ be a vector of fixed functions. For a convex and twice differentiable function $f : (0, \infty) \to \mathbb{R}$, define*

$$\begin{aligned} &F(X_1, \ldots, X_n) \\ &= \sup_{\|w\|_2 = 1, \mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} \left\{ K_f(P_n, P_{\hat{\theta}}; b w^{\mathrm{T}} g_{\mu,\eta_0}) - K_f(P_\epsilon, P_{\hat{\theta}}; b w^{\mathrm{T}} g_{\mu,\eta_0}) \right\} \\ &= \sup_{\|w\|_2 = 1, \mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} \left\{ \frac{1}{n} \sum_{i=1}^n f'(\mathrm{e}^{b w^{\mathrm{T}} g_{\mu,\eta_0}(X_i)}) - \mathrm{E} f'(\mathrm{e}^{b w^{\mathrm{T}} g_{\mu,\eta_0}(X)}) \right\}, \end{aligned}$$

where $g_{\mu,\eta_0}(x) = g(x - \mu) - \eta_0$. Suppose that conditionally on $(X_1, \ldots, X_n)$, the random variable $Z_{n,j} = \sup_{\mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} |n^{-1} \sum_{i=1}^n \epsilon_i g_{\mu,\eta_0,j}(X_i)|$ is sub-gaussian with tail parameter $\sqrt{V_g/n}$ for $j = 1, \ldots, q$, where $(\epsilon_1, \ldots, \epsilon_n)$ are Rademacher variables, independent of $(X_1, \ldots, X_n)$, and $g_{\mu,\eta_0,j} : \mathbb{R}^p \to [-1, 1]$ denotes the jth component of g_{μ,η_0}. Then for any $\delta > 0$, we have that with probability at least $1 - 2\delta$,

$$F(X_1, \ldots, X_n) \leq b R_{2,b\sqrt{q}} \left\{ C_{\mathrm{sg},12} \sqrt{\frac{2qV_g}{n}} + \sqrt{\frac{2q \log(\delta^{-1})}{n}} \right\},$$

where $R_{2,b\sqrt{q}} = \sup_{|u| \leq b\sqrt{q}} \frac{\mathrm{d}}{\mathrm{d}u} f'(\mathrm{e}^u)$ and $C_{\mathrm{sg},12}$ is the universal constant in Lemma 67.

Proof First, F satisfies the bounded difference condition, because $|bw^{\mathrm{T}} g_{\mu,\eta_0}| \le b\sqrt{q}$ with $\|w\|_2 \le 1$ and f' is non-decreasing by the convexity of f:

$$\sup_{X_1,\ldots,X_n,X_i'} \left|F(X_1,\ldots,X_n) - F(X_1,\ldots,X_i',\ldots,X_n)\right| \le \frac{f'(\mathrm{e}^{b\sqrt{q}}) - f'(\mathrm{e}^{-b\sqrt{q}})}{n} \le \frac{2b\sqrt{q}R_{2,b}}{n},$$

where $R_{2,b\sqrt{q}} = \sup_{|u|\le b\sqrt{q}} \frac{\mathrm{d}}{\mathrm{d}u} f'(\mathrm{e}^u)$. By McDiarmid's inequality (McDiarmid, 1989), for any $t > 0$, we have that with probability at least $1 - 2\mathrm{e}^{-2nt^2}$,

$$|F(X_1,\ldots,X_n) - \mathrm{E}F(X_1,\ldots,X_n)| \le 2b\sqrt{q}R_{2,b}t.$$

For any $\delta > 0$, taking $t = \sqrt{\log(\delta^{-1})/(2n)}$ shows that with probability at least $1 - 2\delta$,

$$|F(X_1,\ldots,X_n) - \mathrm{E}F(X_1,\ldots,X_n)| \le bR_{2,b}\sqrt{\frac{2q\log(\delta^{-1})}{n}}.$$

Next, the expectation of $F(X_1,\ldots,X_n)$ can be bounded as follows:

$$\begin{aligned}
&\mathrm{E}\sup_{\|w\|_2=1,\mu\in\mathbb{R}^p,\eta_0\in[0,1]^q}\left\{\frac{1}{n}\sum_{i=1}^n f'(\mathrm{e}^{bw^{\mathrm{T}}g_{\mu,\eta_0}(X_i)}) - \mathrm{E}f'(\mathrm{e}^{bw^{\mathrm{T}}g_{\mu,\eta_0}(X)})\right\} \\
&\le 2\mathrm{E}\sup_{\|w\|_2=1,\mu\in\mathbb{R}^p,\eta_0\in[0,1]^q}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f'(\mathrm{e}^{bw^{\mathrm{T}}g_{\mu,\eta_0}(X_i)})\right\} \qquad (116)\\
&\le 2R_{2,b\sqrt{q}}\,\mathrm{E}\sup_{\|w\|_2=1,\mu\in\mathbb{R}^p,\eta_0\in[0,1]^q}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i bw^{\mathrm{T}}g_{\mu,\eta_0}(X_i)\right\} \qquad (117)\\
&\le 2bR_{2,b\sqrt{q}}\,\mathrm{E}\sup_{\mu\in\mathbb{R}^p,\eta_0\in[0,1]^q}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\mu,\eta_0}(X_i)\right\|_2 \\
&\le 2bR_{2,b\sqrt{q}}C_{\mathrm{sg},12}\sqrt{\frac{2qV_g}{n}}. \qquad (118)
\end{aligned}$$

Line (116) follows from the symmetrization Lemma 77, where $(\epsilon_1,\ldots,\epsilon_n)$ are Rademacher variables, independent of $(X_1,\ldots,X_n)$. Line (117) follows by Lemma 78, because $f'(\mathrm{e}^t)$ is $R_{2,b\sqrt{q}}$-Lipschitz in $u \in [-b\sqrt{q}, b\sqrt{q}]$. Line (118) follows because

$$\begin{aligned}
&\mathrm{E}\sup_{\mu\in\mathbb{R}^p,\eta_0\in[0,1]^q}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\mu,\eta_0}(X_i)\right\|_2 \le \left\{\mathrm{E}\sup_{\mu\in\mathbb{R}^p,\eta_0\in[0,1]^q}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\mu,\eta_0}(X_i)\right\|_2^2\right\}^{1/2} \\
&\le \left\{\sum_{j=1}^q \mathrm{E}\sup_{\mu\in\mathbb{R}^p,\eta_0\in[0,1]^q}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\mu,\eta_0,j}(X_i)\right|^2\right\}^{1/2} \\
&\le C_{\mathrm{sg},12}\sqrt{\frac{2qV_g}{n}}.
\end{aligned}$$

For the last step, we use the assumption that conditionally on $(X_1, \ldots, X_n)$, the random variable $Z_{n,j} = \sup_{\mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} |n^{-1} \sum_{i=1}^n \epsilon_i g_{\mu,\eta_0,j}(X_i)|$ is sub-gaussian with tail parameter $\sqrt{V_g/n}$ for each $j = 1, \ldots, q$, and apply Lemma 67 to obtain $\mathrm{E}(Z_{n,j}^2 | X_1, \ldots, X_n) \leq C_{\mathrm{sg},12}^2 (2V_g/n)$, and then $\mathrm{E}(Z_{n,j}^2) \leq C_{\mathrm{sg},12}^2 (2V_g/n)$.

Combining the tail probability and expectation bounds yields the desired result. ■

Lemma 46 *Suppose that $f : (0, \infty) \to \mathbb{R}$ is convex and three-times differentiable. Let $b > 0$ be fixed.*

(i) For any function $h : \mathbb{R}^p \to [-b, b]$, we have

$$K_f(P_\epsilon, P_{\hat{\theta}}; h) \geq f'(\mathrm{e}^{-b})\epsilon + f'(\mathrm{e}^{\mathrm{E}_{P_{\theta^*}} h(x)}) - f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\hat{\theta}}} h(x)}) - \frac{1}{2} R_{33,b},$$

where $R_{33,b} = R_{31,b} \mathrm{Var}_{P_{\theta^*}} h(x) + R_{32,b} \mathrm{Var}_{P_{\hat{\theta}}} h(x)$, $R_{31,b} = \sup_{|u| \leq b} \frac{\mathrm{d}^2}{\mathrm{d}u^2} \{-f'(\mathrm{e}^u)\}$, and $R_{32,b} = \sup_{|u| \leq b} \frac{\mathrm{d}^2}{\mathrm{d}u^2} f^{\#}(\mathrm{e}^u)$.

(ii) If, in addition, $\mathrm{E}_{P_{\theta^*}} h(x) = 0$ and $\mathrm{E}_{P_{\hat{\theta}}} h(x) \leq 0$, then

$$K_f(P_\epsilon, P_{\hat{\theta}}; h) \geq f'(\mathrm{e}^{-b})\epsilon + R_{4,b} \left\{ \mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\} - \frac{1}{2} R_{33,b},$$

where $R_{4,b} = \inf_{|u| \leq b} \frac{\mathrm{d}}{\mathrm{d}u} f^{\#}(\mathrm{e}^u)$.

Proof (i) First, $K_f(P_\epsilon, P_{\hat{\theta}}; h)$ can be bounded as follows:

$$\begin{aligned}
& K_f(P_\epsilon, P_{\hat{\theta}}; h) \\
&= \epsilon \mathrm{E}_Q f'(\mathrm{e}^{h(x)}) + (1 - \epsilon) \mathrm{E}_{P_{\theta^*}} f'(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_{\hat{\theta}}} f^{\#}(\mathrm{e}^{h(x)}) \\
&\geq f'(\mathrm{e}^{-b})\epsilon + \mathrm{E}_{P_{\theta^*}} f'(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_{\hat{\theta}}} f^{\#}(\mathrm{e}^{h(x)}), \\
&= f'(\mathrm{e}^{-b})\epsilon + K_f(P_{\theta^*}, P_{\hat{\theta}}; h),
\end{aligned}$$

where the inequality follows because $f'(\mathrm{e}^{h(x)}) \geq f'(\mathrm{e}^{-b})$ for $h(x) \in [-b, b]$ by the convexity of f. Next, consider the function

$$\kappa(t) = \mathrm{E}_{P_{\theta^*}} f'(\mathrm{e}^{E_1 + t\tilde{h}_1(x)}) - \mathrm{E}_{P_{\hat{\theta}}} f^{\#}(\mathrm{e}^{E_2 + t\tilde{h}_2(x)}),$$

where $E_1 = \mathrm{E}_{P_{\theta^*}} h(x)$, $E_2 = \mathrm{E}_{P_{\hat{\theta}}} h(x)$, $\tilde{h}_1(x) = h(x) - E_1$, and $\tilde{h}_2(x) = h(x) - E_2$. A Taylor expansion of $\kappa(1) = K_f(P_{\theta^*}, P_{\hat{\theta}}; h)$ about $t = 0$ yields

$$K_f(P_{\theta^*}, P_{\hat{\theta}}; h) = f'(\mathrm{e}^{E_1}) - f^{\#}(\mathrm{e}^{E_2}) - \frac{1}{2} \kappa''(t),$$

where for some $t \in [0, 1]$,

$$\kappa''(t) = -\mathrm{E}_{P_{\theta^*}} \left\{ \tilde{h}_1^2(x) \frac{\mathrm{d}^2}{\mathrm{d}u^2} f'(\mathrm{e}^{E_1 + u})|_{u = t\tilde{h}_1(x)} \right\} + \mathrm{E}_{P_{\hat{\theta}}} \left\{ \tilde{h}_2^2(x) \frac{\mathrm{d}^2}{\mathrm{d}u^2} f^{\#}(\mathrm{e}^{E_2 + u})|_{u = t\tilde{h}_2(x)} \right\}.$$

The desired result then follows because $E_1 + t\tilde{h}_1(x) \in [-b, b]$ and $E_2 + t\tilde{h}_2(x) \in [-b, b]$ for $t \in [0, 1]$ and hence $\kappa''(t) \le R_{33,b}$ by the definition of $R_{33,b}$.

(ii) The inequality from (i) can be rewritten as

$$\begin{aligned} &K_f(P_\epsilon, P_{\hat{\theta}}; h) \\ &\ge f'(\mathrm{e}^{-b})\epsilon + \left\{ f'(\mathrm{e}^{\mathrm{E}_{P_{\theta^*}} h(x)}) - f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\theta^*}} h(x)}) \right\} + f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\theta^*}} h(x)}) - f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\hat{\theta}}} h(x)}) - \frac{1}{2} R_{33,b}. \end{aligned}$$

If $\mathrm{E}_{P_{\theta^*}} h(x) = 0$, then

$$K_f(P_\epsilon, P_{\hat{\theta}}; h) \ge f'(\mathrm{e}^{-b})\epsilon + f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\theta^*}} h(x)}) - f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\hat{\theta}}} h(x)}) - \frac{1}{2} R_{33,b}.$$

Moreover, if $\mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \ge 0$, then

$$f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\theta^*}} h(x)}) - f^{\#}(\mathrm{e}^{\mathrm{E}_{P_{\hat{\theta}}} h(x)}) \ge R_{4,b} \left\{ \mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\}.$$

by the mean value theorem and the definition of $R_{4,b}$. Combining the preceding two displays gives the desired result. ■

Proposition 47 *Let $b_2 > 0$ be fixed and $b_2^\dagger = b_2\sqrt{2p}$. In the setting of Proposition 24, it holds with probability at least $1-2\delta$ that for any $\gamma \in \Gamma_{\mathrm{rp1}}$ with $\mathrm{pen}_2(\gamma) = b_2$, $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0$, and $\mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \le 0$,*

$$\begin{aligned} &K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ &\ge R_{4,b_2^\dagger} \left\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} + f'(\mathrm{e}^{-b_2^\dagger})\epsilon - 4C_{\mathrm{sg},12}^2 M_2 b_2^2 R_{3,b_2^\dagger} - b_2 R_{2,b_2^\dagger} \lambda_{22} \end{aligned}$$

where $R_{3,b} = R_{31,b} + R_{32,b}$ as in Lemma 36 and, with $C_{\mathrm{rad5}} = C_{\mathrm{sg},12} C_{\mathrm{rad3}}$,

$$\lambda_{22} = C_{\mathrm{rad5}} \sqrt{\frac{16p}{n}} + \sqrt{\frac{2p \log(\delta^{-1})}{n}},$$

depending on the universal constants $C_{\mathrm{sg},12}$ and C_{rad3} in Lemma 67 and Corollary 82.

Proof By definition, for any $\gamma \in \Gamma_{\mathrm{rp1}}$, $h_\gamma(x)$ can be represented as $h_{\mathrm{rp1},\beta,c}(x)$ such that $\beta_0 = \gamma_0$ and $\mathrm{pen}_2(\beta) = \sqrt{2}\mathrm{pen}_2(\gamma)$:

$$\begin{aligned} h_{\mathrm{rp1},\beta,c}(x) &= \beta_0 + \sum_{j=1}^{p} \beta_{1j} \mathrm{ramp}(x_j - c_j) \\ &= \beta_0 + \beta_1^{\mathrm{T}} \varphi_{\mathrm{rp},c}(x), \end{aligned}$$

where $c = (c_1, \ldots, c_p)^{\mathrm{T}}$ with $c_j \in \{0, 1\}$, $\beta = (\beta_0, \beta_1^{\mathrm{T}})^{\mathrm{T}}$ with $\beta_1 = (\beta_{11}, \ldots, \beta_{1p})^{\mathrm{T}}$, and $\varphi_{\mathrm{rp},c}(x) : \mathbb{R}^p \to [0, 1]^p$ denotes the vector of functions with the jth component $\mathrm{ramp}(x_j - c_j)$ for $j = 1, \ldots, p$. Then for any $\gamma \in \Gamma_{\mathrm{rp1}}$ with $\mathrm{pen}_2(\gamma) = b_2$, we have $\beta_0 = \gamma_0$ and $\mathrm{pen}_2(\beta) = \sqrt{2}b_2$ correspondingly, and hence $h_\gamma(x) - \gamma_0 = h_{\mathrm{rp1},\beta,c}(x) - \beta_0 \in [-b_2\sqrt{2p}, b_2\sqrt{2p}]$ by the

Cauchy–Schwartz inequality and the boundedness of the ramp function in $[0,1]$. Moreover, $h_{\mathrm{rp1},\beta,c}(x)$ can be expressed in the form $\beta_0 + \mathrm{pen}_2(\beta) w^{\mathrm{T}} g(x)$, where for $q = 2p$, $w \in \mathbb{R}^q$ is an L_2 unit vector, $g : \mathbb{R}^p \to [0,1]^q$ is a vector of functions, including $\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j - 1)$ for $j = 1, \ldots, p$. Parenthetically, at most one of the coefficients in w associated with $\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j - 1)$ is nonzero for each j, although this property is not used in the subsequent discussion.

For any $\gamma \in \Gamma_{\mathrm{rp1}}$ with $\mathrm{pen}_2(\gamma) = b_2$ and $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0$, the function $h_{\gamma,\hat{\mu}}(x)$ can be expressed as

$$h_{\gamma,\hat{\mu}}(x) = \beta_1^{\mathrm{T}} \left\{ \varphi_{\mathrm{rp},c}(x - \hat{\mu}) - \beta_{01} \right\},$$

where $\beta_{01} = \mathrm{E}_{P_{\theta^*}} \varphi_{\mathrm{rp},c}(x - \hat{\mu})$. The mean-centered ramp functions in $\varphi_{\mathrm{rp},c}(x - \hat{\mu}) - \beta_{01}$ are bounded between $[-1, 1]$, and hence $h_{\gamma,\hat{\mu}}(x) \in [-b_2\sqrt{2p}, b_2\sqrt{2p}]$ similarly as above. Moreover, such $h_{\gamma,\hat{\mu}}(x)$ can be expressed in the form $\mathrm{pen}_2(\beta) w^{\mathrm{T}} \{ g(x - \hat{\mu}) - \eta_0 \}$, where $w \in \mathbb{R}^q$ is an L_2 unit vector, $g(x) : \mathbb{R}^p \to [0,1]^q$ is defined as above, and $\eta_0 = \mathrm{E}_{P_{\theta^*}} g(x - \hat{\mu}) \in [0,1]^q$ by the boundedness of the ramp function in $[0,1]$.

Next, $K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ can be bounded as

$$\begin{aligned} & K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ & \geq K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \}. \end{aligned} \tag{119}$$

For any $\gamma \in \Gamma_{\mathrm{rp1}}$ with $\mathrm{pen}_2(\gamma) = b_2$, $E_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0$, and $E_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \leq 0$, applying Lemma 46(ii) with $h = h_{\gamma,\hat{\mu}}$ and $b = b_2^\dagger = b_2\sqrt{2p}$ yields

$$\begin{aligned} & K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \geq f'(\mathrm{e}^{-b_2^\dagger})\epsilon \\ & \quad + R_{4,b_2^\dagger} \left\{ \mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\} - \frac{1}{2} \left\{ R_{31,b_2^\dagger} \mathrm{Var}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) + R_{32,b_2^\dagger} \mathrm{Var}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\}. \end{aligned}$$

By Lemma 62(i), $\mathrm{Var}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x)$ can bounded as follows:

$$\begin{aligned} & \mathrm{Var}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = \mathrm{Var}_{P_{\theta^*}} \beta_1^{\mathrm{T}} \varphi_{\mathrm{rp},c}(x - \hat{\mu}) \\ & \leq \|\beta_1\|_2^2 \cdot 2 C_{\mathrm{sg},12}^2 (\sqrt{2})^2 \|\Sigma^*\|_{\mathrm{op}} = 4\mathrm{pen}_2^2(\beta) C_{\mathrm{sg},12}^2 M_2. \end{aligned}$$

Similarly, $\mathrm{Var}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x)$ can also be bounded by $4\mathrm{pen}_2^2(\beta) C_{\mathrm{sg},12}^2 M_2$, because $\|\hat{\Sigma}\|_{\mathrm{op}} \leq M_2$. Hence $K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ can be bounded as

$$\begin{aligned} & K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ & \geq f'(\mathrm{e}^{-b_2^\dagger})\epsilon + R_{4,b_2^\dagger} \left\{ \mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\} - 4 C_{\mathrm{sg},12}^2 M_2 b_2^2 R_{3,b_2^\dagger}, \end{aligned} \tag{120}$$

where $R_{3,b} = R_{31,b} + R_{32,b}$ as in Lemma 36. Moreover, by Lemma 34 with $b = \sqrt{2} b_2$ and $g(x) \in [0,1]^q$ defined above, it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{\mathrm{rp1}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_2(\gamma) = b_2$,

$$\begin{aligned} & \{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \} \\ & \leq b_2 R_{2,b_2\sqrt{q}} \left\{ C_{\mathrm{sg},12} \sqrt{\frac{2qV_g}{n}} + \sqrt{\frac{q \log(\delta^{-1})}{n}} \right\} \\ & \leq b_2 R_{2,b_2\sqrt{2p}} \left\{ C_{\mathrm{sg},12} C_{\mathrm{rad}3} \sqrt{\frac{16p}{n}} + \sqrt{\frac{2p \log(\delta^{-1})}{n}} \right\}, \end{aligned} \tag{121}$$

where $V_g = 4C_{\mathrm{rad3}}^2$ is determined in Lemma 34 as follows. For $j = 1, \ldots, q$, consider the function class $\mathcal{G}_j = \{g_{\mu,\eta_0,j} : \mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q\}$, where $\mu = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$, $\eta_0 = (\eta_{01}, \ldots, \eta_{0q})^{\mathrm{T}}$, and, as defined in Lemma 34, $g_{\mu,\eta_0,j}(x)$ is of the form $\mathrm{ramp}(x_{j_1} - \mu_{j_1}) - \eta_{0j}$ or $\mathrm{ramp}(x_{j_1} - \mu_{j_1} - 1) - \eta_{0j}$. By Lemma 60, the VC index of moving-knot ramp functions is 2. By Lemma 61, the VC index of constant functions is also 2. By applying Corollary 82 (ii) with vanishing $\mathcal{G}$, we obtain that conditionally on $(X_1, \ldots, X_n)$, the random variable $Z_{n,j} = \sup_{\mu\in\mathbb{R}^p,\eta_0\in[0,1]^q} |n^{-1}\sum_{i=1}^n \epsilon_i g_{\mu,\eta_0,j}(X_i)| = \sup_{f_j\in\mathcal{G}_j} |n^{-1}\sum_{i=1}^n \epsilon_i f_j(X_i)|$ is sub-gaussian with tail parameter $C_{\mathrm{rad3}}\sqrt{4/n}$ for $j = 1, \ldots, q$.

Combining the inequalities (119)–(121) leads to the desired result. ■

Proposition 48 *In the setting of Proposition 24 or Proposition 26, suppose that for $a \in (0, 1/2)$,*

$$D \stackrel{def}{=} \sup_{\gamma\in\Gamma_{\mathrm{rp1}},\mathrm{pen}_2(\gamma)=\sqrt{1/2}} \left\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} \le a. \tag{122}$$

Then we have

$$\|\hat{\mu} - \mu^*\|_2 \le S_{4,a} D, \tag{123}$$

$$\|\hat{\sigma} - \sigma^*\|_2 \le S_5 \left(D + \|\hat{\mu} - \mu^*\|_2/2\right) \le S_{6,a} D, \tag{124}$$

where $S_{4,a}$, S_5, and $S_{6,a}$ are defined as in Lemma 38 and Remark 39 with $M = M_2$.

Proof For any $\gamma \in \Gamma_{\mathrm{rp1}}$ with $\mathrm{pen}_2(\gamma) = \sqrt{1/2}$, the function $h_\gamma(x)$ can be obtained as $h_{\mathrm{rp1},\beta,c}(x)$ with $\mathrm{pen}_2(\beta) = 1$. For $j = 1, \ldots, p$, we restrict $h_{\gamma,\hat{\mu}}(x)$ in (122) such that $h_\gamma(x)$ is a ramp function of x_j, in the form $\pm\mathrm{ramp}(x_j - c)$ for $c \in \{0, 1\}$. Applying Lemma 38 shows that there exists $h_j^{(1)}(x_j)$ in the form $\pm\mathrm{ramp}(x_j)$ and $h_j^{(2)}(x_j)$ in the form $\pm\mathrm{ramp}(x_j - 1)$ such that

$$|\hat{\mu}_j - \mu_j^*| \le S_{4,a} \left\{ \mathrm{E}_{P_{\theta^*}} h_j^{(1)}(x_j - \hat{\mu}_j) - \mathrm{E}_{P_{\hat{\theta}}} h_j^{(1)}(x_j - \hat{\mu}_j) \right\}, \tag{125}$$

$$|\hat{\sigma}_j - \sigma_j^*| \le S_5 \left\{ \mathrm{E}_{P_{\theta^*}} h_j^{(2)}(x_j - \hat{\mu}_j) - \mathrm{E}_{P_{\hat{\theta}}} h_j^{(2)}(x_j - \hat{\mu}_j) \right\} + S_5 |\hat{\mu}_j - \mu_j^*|/2. \tag{126}$$

From (125), we have that for any L_2 unit vector $w = (w_1, \ldots, w_p)^{\mathrm{T}}$,

$$\begin{aligned} \sum_{j=1}^p |w_j(\hat{\mu}_j - \mu_j)| &\le S_{4,a} \sum_{j=1}^p |w_j| \left\{ \mathrm{E}_{P_{\theta^*}} h_j^{(1)}(x_j - \hat{\mu}_j) - \mathrm{E}_{P_{\hat{\theta}}} h_j^{(1)}(x_j - \hat{\mu}_j) \right\} \\ &= S_{4,a} \left\{ \mathrm{E}_{P_{\theta^*}} h^{(1)}(x - \hat{\mu}) - \mathrm{E}_{P_{\hat{\theta}}} h^{(1)}(x - \hat{\mu}) \right\}, \end{aligned}$$

where $h^{(1)}(x) = \sum_{j=1}^p |w_j| h_j^{(1)}(x_j)$. In fact, $h^{(1)}(x)$ can be expressed as $h_{\mathrm{rp1},\beta,c}(x)$ such that $c = (0, \ldots, 0)^{\mathrm{T}}$ and each component in β_1 is either $|w_j|$ or $-|w_j|$ for $j = 1, \ldots, p$, which implies that $\mathrm{pen}_2(\beta) = \|w\|_2 = 1$. Hence by the definition of D, we obtain (123).

Similarly, from (126), we have that for any L_2 unit vector $w = (w_1, \ldots, w_p)^{\mathrm{T}}$,

$$\begin{aligned}
&\sum_{j=1}^{p} |w_j(\hat{\sigma}_j - \sigma_j)| \\
&\leq S_5 \sum_{j=1}^{p} |w_j| \left\{ \mathrm{E}_{P_{\theta^*}} h_j^{(2)}(x_j - \hat{\mu}_j) - \mathrm{E}_{P_{\hat{\theta}}} h_j^{(2)}(x_j - \hat{\mu}_j) \right\} + S_5 \sum_{j=1}^{p} |w_j(\hat{\mu}_j - \mu_j^*)|/2 \\
&= S_5 \left\{ \mathrm{E}_{P_{\theta^*}} h^{(2)}(x - \hat{\mu}) - \mathrm{E}_{P_{\hat{\theta}}} h^{(2)}(x - \hat{\mu}) \right\} + S_5 |w^{\mathrm{T}}(\hat{\mu} - \mu^*)|/2,
\end{aligned}$$

where $h^{(2)}(x) = \sum_{j=1}^{p} |w_j| h_j^{(2)}(x_j)$, which can be expressed in the form $h_{\mathrm{rp1},\beta,c}(x)$ with $c = (1, \ldots, 1)^{\mathrm{T}}$ and $\mathrm{pen}_2(\beta) = \|w\|_2 = 1$. Hence by the definition of D, we obtain (124). ■

Proposition 49 *Let $b_3 > 0$ be fixed and $b_3^\dagger = b_3\sqrt{p(p-1)}$. In the setting of Proposition 24, it holds with probability at least $1-2\delta$ that for any $\gamma \in \Gamma_{\mathrm{rp2}}$ with $\mathrm{pen}_2(\gamma) = b_3$, $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0$, and $\mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \leq 0$,*

$$\begin{aligned}
&K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\
&\geq R_{4,2b_3^\dagger} \left\{ \mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\} + f'(\mathrm{e}^{-2b_3^\dagger})\epsilon - (80 C_{\mathrm{sg},12}^2 M_2) p b_3^2 R_{3,2b_3^\dagger} - \sqrt{p} b_3 R_{2,b_3^\dagger} \lambda_{32},
\end{aligned}$$

where $R_{3,b} = R_{31,b} + R_{32,b}$ as in Lemma 36 and, with $C_{\mathrm{rad5}} = C_{\mathrm{sg},12} C_{\mathrm{rad3}}$,

$$\lambda_{32} = C_{\mathrm{rad5}} \sqrt{\frac{12(p-1)}{n}} + \sqrt{\frac{(p-1)\log(\delta^{-1})}{n}},$$

depending on the universal constants $C_{\mathrm{sg},12}$ and C_{rad3} in Lemma 67 and Corollary 82.

Proof By definition, for any $\gamma \in \Gamma_{\mathrm{rp2}}$, $h_\gamma(x)$ can be represented as $h_{\mathrm{rp2},\beta}(x)$ such that $\beta_0 = \gamma_0$ and $\mathrm{pen}_2(\beta) = 2\mathrm{pen}_2(\gamma)$, where

$$\begin{aligned}
h_{\mathrm{rp2},\beta}(x) &= \beta_0 + \sum_{1 \leq i \neq j \leq p} \beta_{2,ij} \mathrm{ramp}(x_i) \mathrm{ramp}(x_j) \\
&= \beta_0 + \beta_2^{\mathrm{T}} \mathrm{vec}(\varphi_{\mathrm{rp}}(x) \otimes \varphi_{\mathrm{rp}}(x)),
\end{aligned}$$

where $\beta = (\beta_0, \beta_2^{\mathrm{T}})^{\mathrm{T}}$ with $\beta_2 = (\beta_{2,ij} : 1 \leq i \neq j \leq p)^{\mathrm{T}}$, and $\varphi_{\mathrm{rp}}(x) : \mathbb{R}^p \to [0,1]^p$ denotes the vector of functions with the jth component $\mathrm{ramp}(x_j)$ for $j = 1, \ldots, p$. Then for any $\gamma \in \Gamma_{\mathrm{rp2}}$ with $\mathrm{pen}_2(\gamma) = b_3$, we have $\beta_0 = \gamma_0$ and $\mathrm{pen}_2(\beta) = 2b_3$ correspondingly, and hence $h_\gamma(x) - \gamma_0 = h_{\mathrm{rp2},\beta}(x) - \beta_0 \in [-2b_3\sqrt{p(p-1)}, 2b_3\sqrt{p(p-1)}]$, by the boundedness of the ramp function in $[0,1]$ and the Cauchy–Schwartz inequality, $\|\beta_2\|_1 \leq \sqrt{p(p-1)}\|\beta_2\|_2$. Moreover, $h_{\mathrm{rp2},\beta}(x)$ can be expressed in the form $\beta_0 + \mathrm{pen}_2(\beta) w^{\mathrm{T}} g(x)$, where for $q = p(p-1)$, $w \in \mathbb{R}^q$ is an L_2 unit vector, $g : \mathbb{R}^p \to [0,1]^q$ is a vector of functions, including $\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$ for $1 \leq i \neq j \leq p$. For symmetry, $\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j)\mathrm{ramp}(x_i)$ are included as two distinct components in g, and the corresponding coefficients are assumed to be identical to each other in w.

For any $\gamma \in \Gamma_{\mathrm{rp2}}$ with $\mathrm{pen}_2(\gamma) = b_3$ and $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0$, the function $h_{\gamma,\hat{\mu}}(x)$ can be expressed as

$$h_{\gamma,\hat{\mu}}(x) = \beta_2^{\mathrm{T}} \left\{ \varphi_{\mathrm{rp}}(x - \hat{\mu}) - \beta_0 \right\},$$

where $\beta_0 = \mathrm{E}_{P_{\theta^*}} \varphi_{\mathrm{rp}}(x - \hat{\mu})$. The mean-centered ramp functions in $\varphi_{\mathrm{rp}}(x - \hat{\mu}) - \beta_0$ are bounded between $[-1, 1]$, and hence $h_{\gamma,\hat{\mu}}(x) \in [-b_3 2p, b_3 2p]$ similarly as above. Moreover, such $h_{\gamma,\hat{\mu}}(x)$ can be expressed in the form $\mathrm{pen}_2(\beta) w^{\mathrm{T}} \{ g(x - \hat{\mu}) - \eta_0 \}$, where $w \in \mathbb{R}^q$ is an L_2 unit vector, $g(x) : \mathbb{R}^p \to [0,1]^q$ is defined as above, and $\eta_0 = \mathrm{E}_{P_{\theta^*}} g(x - \hat{\mu}) \in [0,1]^q$ by the boundedness of the ramp function in $[0, 1]$.

Next, $K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ can be bounded as

$$\begin{aligned} &K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ &\geq K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - |K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})|. \end{aligned} \tag{127}$$

For any $\gamma \in \Gamma_{\mathrm{rp2}}$ with $\mathrm{pen}_2(\gamma) = b_3$, $E_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0$, and $E_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \leq 0$, applying Lemma 46(ii) with $h = h_{\gamma,\hat{\mu}}$ and $b = 2b_3^\dagger = 2b_3\sqrt{p(p-1)}$ yields

$$\begin{aligned} &K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \geq f'(\mathrm{e}^{-2b_3^\dagger})\epsilon \\ &\quad + R_{4,2b_3^\dagger} \left\{ \mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\} - \frac{1}{2} \left\{ R_{31,2b_3^\dagger} \mathrm{Var}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) + R_{32,2b_3^\dagger} \mathrm{Var}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\}. \end{aligned}$$

By Lemma 63(ii), $\mathrm{Var}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x)$ can bounded as follows:

$$\begin{aligned} &\mathrm{Var}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) \\ &= \mathrm{Var}_{P_{\theta^*}} \beta_2^{\mathrm{T}} \mathrm{vec}(\varphi_{\mathrm{rp}}(x - \hat{\mu}) \otimes \varphi_{\mathrm{rp}}(x - \hat{\mu})) \\ &\leq \|\beta_2\|_2^2 \cdot 20 C_{\mathrm{sg},12}^2 (\sqrt{2})^2 p \|\Sigma^*\|_{\mathrm{op}} \\ &\leq 40 \mathrm{pen}_2^2(\beta) C_{\mathrm{sg},12}^2 p M_2. \end{aligned}$$

Similarly, $\mathrm{Var}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x)$ can also be bounded by $40\mathrm{pen}_2^2(\beta) C_{\mathrm{sg},12}^2 p M_2$, because $\|\hat{\Sigma}\|_{\mathrm{op}} \leq M_2$. Hence, with $\mathrm{pen}_2(\beta) = 2b_3$, $K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ can be bounded as

$$\begin{aligned} &K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ &\geq f'(\mathrm{e}^{-2b_3^\dagger})\epsilon + R_{4,2b_3^\dagger} \left\{ \mathrm{E}_{P_{\theta^*}} h(x) - \mathrm{E}_{P_{\hat{\theta}}} h(x) \right\} - 80 C_{\mathrm{sg},12}^2 p M_2 b_3^2 R_{3,2b_3^\dagger}, \end{aligned} \tag{128}$$

where $R_{3,b} = R_{31,b} + R_{32,b}$ as in Lemma 36. Moreover, by Lemma 34 with $b = 2b_3$ and $g(x) \in [0,1]^q$ defined above, it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{\mathrm{rp2}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_2(\gamma) = b_3$,

$$\begin{aligned} &|K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_f(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})| \\ &\leq b_3 R_{2,b_3\sqrt{q}} \left\{ C_{\mathrm{sg},12} \sqrt{\frac{2qV_g}{n}} + \sqrt{\frac{q \log(\delta^{-1})}{n}} \right\} \\ &= b_3 R_{2,b_3^\dagger} \left\{ C_{\mathrm{sg},12} C_{\mathrm{rad}3} \sqrt{\frac{12p(p-1)}{n}} + \sqrt{\frac{p(p-1)\log(\delta^{-1})}{n}} \right\}, \end{aligned} \tag{129}$$

where $V_g = 6C_{\mathrm{rad3}}^2$ is determined in Lemma 34 as follows. For $j = 1, \ldots, q$, consider the function class $\mathcal{G}_j = \{g_{\mu,\eta_0,j} : \mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q\}$, where $\mu = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$, $\eta_0 = (\eta_{01}, \ldots, \eta_{0q})^{\mathrm{T}}$, and, as defined in Lemma 34, $g_{\mu,\eta_0,j}(x)$ is of the form $\mathrm{ramp}(x_{j_1} - \mu_{j_1})\mathrm{ramp}(x_{j_2} - \mu_{j_2}) - \eta_{0j}$ for $1 \le j_1 \ne j_2 \le p$. By Lemma 60, the VC index of moving-knot ramp functions is 2. By Lemma 61, the VC index of constant functions is also 2. By applying Corollary 82 (ii), we obtain that conditionally on $(X_1, \ldots, X_n)$, the random variable $Z_{n,j} = \sup_{\mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} |n^{-1} \sum_{i=1}^n \epsilon_i g_{\mu,\eta_0,j}(X_i)| = \sup_{f_j \in \mathcal{G}_j} |n^{-1} \sum_{i=1}^n \epsilon_i f_j(X_i)|$ is subgaussian with tail parameter $C_{\mathrm{rad3}}\sqrt{6/n}$ for $j = 1, \ldots, q$.

Combining the inequalities (127)–(129) leads to the desired result. ■

Proposition 50 *In the setting of Proposition 24 or Proposition 26, denote*

$$D = \sup_{\gamma \in \Gamma_{\mathrm{rp2}}, \mathrm{pen}_2(\gamma) = 1/2} \left\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\}.$$

Then we have

$$\|\hat{\Sigma} - \Sigma^*\|_{\mathrm{F}} \le 2M_2^{1/2}\sqrt{p}\|\hat{\sigma} - \sigma^*\|_2 + S_7(\sqrt{2p}\Delta_{\hat{\mu},\hat{\sigma}} + D),$$

where $\Delta_{\hat{\mu},\hat{\sigma}} = (\|\hat{\mu} - \mu^*\|_2^2 + \|\hat{\sigma} - \sigma^*\|_2^2)^{1/2}$ *and* S_7 *is defined as in Lemma 40 with* $M = M_2$.

Proof For any $\gamma \in \Gamma_{\mathrm{rp2}}$ with $\mathrm{pen}_2(\gamma) = 1/2$, the function $h_\gamma(x) \in \mathcal{H}_{\mathrm{rp2}}$ can be obtained as $h_{\mathrm{rp2},\beta}(x)$ with $\mathrm{pen}_2(\beta) = 1$. First, we handle the effect of different means and standard deviations between P_{θ^*} and $P_{\hat{\theta}}$ in D. Denote

$$D^\dagger = \sup_{\gamma \in \Gamma_{\mathrm{rp2}}, \mathrm{pen}_2(\gamma) = 1/2} \left\{ \mathrm{E}_{P_{\theta^\dagger}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\},$$

where $\theta^\dagger = (\hat{\mu}, \mathrm{diag}(\hat{\sigma})\Sigma_0^*\mathrm{diag}(\hat{\sigma}))$ and Σ_0^* is defined as the correlation matrix such that $\Sigma^* = \mathrm{diag}(\sigma^*)\Sigma_0^*\mathrm{diag}(\sigma^*)$. Then $D^\dagger$ can be related to D as follows:

$$D^\dagger \le D + \sqrt{2p}\Delta_{\hat{\mu},\hat{\sigma}}, \tag{130}$$

where $\Delta_{\hat{\mu},\hat{\sigma}} = (\|\hat{\mu} - \mu^*\|_2^2 + \|\hat{\sigma} - \sigma^*\|_2^2)^{1/2}$. In fact, by Lemma 65 with $g(x)$ set to $\varphi_{\mathrm{rp}}(x)$, which is $(1/2)$-Lipschitz and componentwise bounded in $[0,1]$, we have that for any $\gamma \in \Gamma_{\mathrm{rp2}}$ with $\mathrm{pen}_2(\gamma) = 1/2$,

$$\left| \mathrm{E}_{P_{\theta^\dagger}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) \right| \le \sqrt{2p}\Delta_{\hat{\mu},\hat{\sigma}}.$$

For each pair $1 \le i \ne j \le p$, we restrict $h_{\gamma,\hat{\mu}}(x)$ such that $h_\gamma(x)$ is $\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$ or $-\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$. Applying Lemma 40 shows that there exists $r_{ij} \in \{-1, 1\}$ such that

$$|\hat{\rho}_{ij} - \rho_{ij}^*|\hat{\sigma}_i\hat{\sigma}_j \le S_7 r_{ij} \left\{ \mathrm{E}_{P_{\theta^\dagger}} h_{ij}(x - \hat{\mu}) - \mathrm{E}_{P_{\hat{\theta}}} h_{ij}(x - \hat{\mu}) \right\},$$

where $h_{ij}(x) = \mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$. By the triangle inequality, we have

$$\begin{aligned}
&|\hat{\rho}_{ij}\hat{\sigma}_i\hat{\sigma}_j - \rho^*_{ij}\sigma^*_i\sigma^*_j| \\
&\le \hat{\sigma}_i|\hat{\sigma}_j - \sigma^*_j| + \sigma^*_j|\hat{\sigma}^*_i - \sigma^*_i| + |\hat{\rho}_{ij} - \rho^*_{ij}|\hat{\sigma}_i\hat{\sigma}_j \\
&\le M_2^{1/2}|\hat{\sigma}_i - \sigma^*_i| + M_2^{1/2}|\hat{\sigma}_j - \sigma^*_j| + S_7 r_{ij}\left\{\mathrm{E}_{P_{\theta^\dagger}} h_{ij}(x-\hat{\mu}) - \mathrm{E}_{P_{\hat{\theta}}} h_{ij}(x-\hat{\mu})\right\}.
\end{aligned}$$

In addition, we have $|\hat{\sigma}_i^2 - \sigma_i^{*2}| = |(\hat{\sigma}_i + \sigma^*_i)(\hat{\sigma}_i - \sigma^*_i)| \le 2M_2^{1/2}|\hat{\sigma}_i - \sigma^*_i|$. Then for any L_2 unit vector $w = (w_{ij} : 1 \le i, j \le p)^{\mathrm{T}} \in \mathbb{R}^{p\times p}$,

$$\begin{aligned}
&\sum_{i=1}^p \left|w_{ii}(\hat{\sigma}_i^2 - \sigma_i^{*2})\right| + \sum_{1\le i\ne j\le p} \left|w_{ij}(\hat{\rho}_{ij}\hat{\sigma}_i\hat{\sigma}_j - \rho^*_{ij}\sigma^*_1\sigma^*_2)\right| \\
&\le 2M_2^{1/2}\sum_{i=1}^p |w_{ii}||\hat{\sigma}_i - \sigma^*_i| + M_2^{1/2}\sum_{1\le i\ne j\le p} |w_{ij}|\left(|\hat{\sigma}_i - \sigma^*_i| + |\hat{\sigma}_j - \sigma^*_j|\right) \\
&\quad + S_7 \sum_{1\le i\ne j\le p} |w_{ij}| r_{ij}\left\{\mathrm{E}_{P_{\theta^\dagger}} h_{ij}(x-\hat{\mu}) - \mathrm{E}_{P_{\hat{\theta}}} h_{ij}(x-\hat{\mu})\right\} \\
&= M_2^{1/2}\sum_{1\le i,j\le p} |w_{ij}|\left(|\hat{\sigma}_i - \sigma^*_i| + |\hat{\sigma}_j - \sigma^*_j|\right) + S_7\left\{\mathrm{E}_{P_{\theta^\dagger}} h(x-\hat{\mu}) - \mathrm{E}_{P_{\hat{\theta}}} h(x-\hat{\mu})\right\},
\end{aligned}$$

where $h(x) = \sum_{1\le i\ne j\le p} |w_{ij}| r_{ij} h_{ij}(x)$. The function $h(x)$ can be expressed as $h_{\mathrm{rp2},\beta}(x)$ such that $\beta_{2,ii} = 0$ for $i = 1, \ldots, p$ and $\beta_{2,ij} = |w_{ij}| r_{ij}$ for $1 \le i \ne j \le p$, and hence $\mathrm{pen}_2(\beta) \le \|w\|_2 = 1$. By the definition of $D^\dagger$, we have $\mathrm{E}_{P_{\theta^\dagger}} h(x-\hat{\mu}) - \mathrm{E}_{P_{\hat{\theta}}} h(x-\hat{\mu}) \le D^\dagger$. Moreover, by the Cauchy–Schwartz inequality, $\sum_{1\le i,j\le p} |w_{ij}||\hat{\sigma}_i - \sigma^*_i| \le \sqrt{p}\|\hat{\sigma} - \sigma^*\|_2$. Substituting these inequalities into the preceding display shows that

$$\begin{aligned}
&\sum_{i=1}^p \left|w_{ii}(\hat{\sigma}_i^2 - \sigma_i^{*2})\right| + \sum_{1\le i\ne j\le p} \left|w_{ij}(\hat{\rho}_{ij}\hat{\sigma}_i\hat{\sigma}_j - \rho^*_{ij}\sigma^*_1\sigma^*_2)\right| \\
&\le 2M_2^{1/2}\sqrt{p}\|\hat{\sigma} - \sigma^*\|_2 + S_7 D^\dagger. \qquad (131)
\end{aligned}$$

Combining (130) and (131) yields the desired result. ■

C.4 Details in main proof of Theorem 15

For completeness, we restate Proposition 21 in the main proof of Theorem 15 below. For $\delta \in (0, 1/7)$, define

$$\begin{aligned}
\lambda_{11} &= \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{n}} + \frac{2\log(5p) + \log(\delta^{-1})}{n}, \\
\lambda_{12} &= 2C_{\mathrm{rad}4}\sqrt{\frac{\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}},
\end{aligned}$$

where $C_{\mathrm{rad}4} = C_{\mathrm{sg}6}C_{\mathrm{rad}3}$, depending on universal constants $C_{\mathrm{sg}6}$ and $C_{\mathrm{rad}3}$ in Lemmas 70 and Corollary 82 in Appendix E. Denote

$$\mathrm{Err}_{h1}(n, p, \delta, \epsilon) = 3\epsilon + 2\sqrt{\epsilon/(n\delta)} + \lambda_{12} + \lambda_1.$$

Proposition 21 (restated) *Assume that* $\|\Sigma_*\|_{\max} \le M_1$ *and* $\epsilon \le 1/5$. *Let* $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ *be a solution to (20) with* $\lambda_1 \ge C_{\text{sp13}} M_{11} \lambda_{11}$ *where* $M_{11} = M_1^{1/2}(M_1^{1/2} + 2\sqrt{2\pi})$ *and* $C_{\text{sp13}} = (5/3)(C_{\text{sp11}} \vee C_{\text{sp12}})$, *depending on universal constants* C_{sp11} *and* C_{sp12} *in Lemma 30 in Appendix C. If* $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \le 1/5$ *and* $\text{Err}_{h1}(n, p, \delta, \epsilon) \le a$ *for a constant* $a \in (0, 1/2)$, *then the following holds with probability at least* $1 - 7\delta$ *uniformly over contamination distribution* Q,

$$\|\hat{\mu} - \mu^*\|_\infty \le S_{4,a} \text{Err}_{h1}(n, p, \delta, \epsilon),$$
$$\|\hat{\Sigma} - \Sigma^*\|_{\max} \le S_{8,a} \text{Err}_{h1}(n, p, \delta, \epsilon),$$

where $S_{4,a} = (1 + \sqrt{2M_1 \log \frac{2}{1-2a}})/a$ *and* $S_{8,a} = 2M_1^{1/2} S_{6,a} + S_7(1 + S_{4,a} + S_{6,a})$ *with* $S_{6,a} = S_5(1 + S_{4,a}/2)$, $S_5 = 2\sqrt{2\pi}(1 - \mathrm{e}^{-2/M_1})^{-1}$, *and* $S_7 = 8\pi M_1 \mathrm{e}^{1/(4M_1)}$ $S_7 = 4\{(\frac{1}{\sqrt{2\pi M_1}} \mathrm{e}^{-1/(8M_1)}) \vee (1 - 2\mathrm{e}^{-1/(8M_1)})\}^{-2}$.

To see why Proposition 21 leads to Theorem 15, we show that conditions in Proposition 21 are satisfied under the setting of Theorem 15. For a constant $a \in (0, 1/2)$, let

$$C_1 = 2\sqrt{3} C_{\text{sp13}} M_{11},$$
$$C_2 = \frac{1}{5} \wedge \frac{\sqrt{3} C_1}{6} \wedge \left(3 \vee \left(\frac{\sqrt{2} C_1}{4C_{\text{rad4}} + 2} + 1\right)\right),$$
$$C = S_{8,a} \left(3 \vee \left(\frac{\sqrt{2} C_1}{4C_{\text{rad4}} + 2} + 1\right)\right).$$

Then the following conditions

(i) $\lambda_1 \ge C_1 \left(\sqrt{\log p/n} + \sqrt{\log(1/\delta)/n}\right)$,

(ii) $\epsilon + \sqrt{\epsilon/(n\delta)} + \lambda_1 \le C_2$,

imply the conditions

(iii) $\lambda_1 \ge C_{\text{sp13}} M_{11} \lambda_{11}$,

(iv) $\text{Err}_{h1}(n, p, \delta, \epsilon) \le a$,

(v) $\epsilon \le 1/5$ and $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \le 1/5$.

In fact, condition (v) follows directly from condition (ii) because $C_2 \le 1/5$. For conditions (iii) and (iv), we first show that λ_{11} and λ_{12} can be upper bounded as follows:

$$\lambda_{11} \le \sqrt{\frac{2\log p + 3\log(\delta^{-1})}{n}} + \frac{2\log p + 3\log(\delta^{-1})}{n} \tag{132}$$
$$\le 2\sqrt{\frac{2\log p + 3\log(\delta^{-1})}{n}} \le 2\sqrt{\frac{2\log p}{n}} + 2\sqrt{\frac{3\log(\delta^{-1})}{n}}, \tag{133}$$

and

$$\lambda_{12} \le 2C_{\mathrm{rad4}}\sqrt{\frac{2\log 2 + 2\log p}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}} \tag{134}$$

$$\le 2C_{\mathrm{rad4}}\sqrt{\frac{2\log p}{n}} + (2C_{\mathrm{rad4}}+1)\sqrt{\frac{2\log(\delta^{-1})}{n}}. \tag{135}$$

Lines (132) and (135) hold because $\log(1/\delta) \ge \log(5)$ for $\delta \in (0, 1/7)$. Line (133) holds because $\sqrt{\frac{2\log p + 3\log(\delta^{-1})}{n}} \le 1$ and hence the linear term in λ_{11} is upper bounded by the square root term. To see this, by conditions (i) and (ii) we have

$$\sqrt{\frac{2\log p + 3\log(\delta^{-1})}{n}} \le \sqrt{\frac{3\log p}{n}} + \sqrt{\frac{3\log(\delta^{-1})}{n}} \le \frac{\sqrt{3}\lambda_1}{C_1} \le \frac{\sqrt{3}C_2}{C_1} \le 1.$$

Line (134) holds because $\log\left(2p(p+1)\right) \le 2\log 2 + 2\log p$ for $p \ge 1$. With the above upper bounds for λ_{11} and λ_{12}, we show that condition (i) implies condition (iii) as follows:

$$\begin{aligned}\lambda_1 &\ge C_1\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)\\ &= 2\sqrt{3}C_{\mathrm{sp13}}M_{11}\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) \ge C_{\mathrm{sp13}}M_{11}\lambda_{11},\end{aligned}$$

and condition (ii) implies condition (iv) as follows:

$$\begin{aligned}\mathrm{Err}_{h1}(n,p,\delta,\epsilon) &= 3\epsilon + 2\sqrt{\epsilon/(n\delta)} + \lambda_{12} + \lambda_1\\ &\le 3\left(\epsilon + \sqrt{\frac{\epsilon}{n\delta}}\right) + \left(\frac{\sqrt{2}C_1}{4C_{\mathrm{rad4}}+2} + 1\right)\lambda_1\\ &\le \left(3 \vee \left(\frac{\sqrt{2}C_1}{4C_{\mathrm{rad4}}+2} + 1\right)\right)\left(\epsilon + \sqrt{\frac{\epsilon}{n\delta}} + \lambda_1\right)\\ &\le \left(3 \vee \left(\frac{\sqrt{2}C_1}{4C_{\mathrm{rad4}}+2} + 1\right)\right)C_2 \le a.\end{aligned}$$

Therefore, Proposition 21 implies Theorem 15 with constant $C = S_{8,a}\left(3 \vee \left(\frac{\sqrt{2}C_1}{4C_{\mathrm{rad4}}+2} + 1\right)\right)$.

Lemma 51 *Consider the hinge GAN (3).*

(i) For any $\epsilon \in [0,1]$ and any function $h : \mathbb{R}^p \to \mathbb{R}$, we have

$$K_{\mathrm{HG}}(P_\epsilon, P_{\theta^*}; h) \le 2\epsilon.$$

(ii) For any function $h : \mathbb{R}^p \to \mathbb{R}$, we have

$$K_{\mathrm{HG}}(P_n, P_{\theta^*}; h) \le 2\hat{\epsilon} + |\mathrm{E}_{P_{\theta^*,n}}h(x) - \mathrm{E}_{P_{\theta^*}}h(x)|, \tag{136}$$

where $\hat{\epsilon} = n^{-1}\sum_{i=1}^n U_i$ and $P_{\theta^*,n}$ denotes the empirical distribution of $\{X_i : U_i = 0, i = 1,\ldots,n\}$ in the latent representation of Huber's contamination model.

Proof (i) For any $h : \mathbb{R}^p \to \mathbb{R}$ we have

$$\begin{aligned}
&K_{\mathrm{HG}}(P_\epsilon, P_{\theta^*}; h) \\
&= \epsilon \mathrm{E}_Q \min(1, h(x)) + (1-\epsilon)\mathrm{E}_{P_{\theta^*}} \min(1, h(x)) + \mathrm{E}_{P_{\theta^*}} \min(1, -h(x)) \\
&\le \epsilon + (1-\epsilon)\mathrm{E}_{P_{\theta^*}} \min(1, h(x)) + \mathrm{E}_{P_{\theta^*}} \min(1, -h(x)) && (137) \\
&\le 2\epsilon + (1-\epsilon)\left\{\mathrm{E}_{P_{\theta^*}} \min(1, h(x)) + \mathrm{E}_{P_{\theta^*}} \min(1, -h(x))\right\} && (138) \\
&\le 2\epsilon. && (139)
\end{aligned}$$

Inequalities (137) and (138) hold because $\min(1, u) \vee \min(1, -u) \le 1$ for all $u \in \mathbb{R}$. Inequality (139) holds because $\min(1, u) + \min(1, -u) \le 0$ for all $u \in \mathbb{R}$.

(ii) Because both $\min(1, u)$ and $\min(1, -u)$ are concave in $u \in \mathbb{R}$ and upper bounded by 1, we have

$$\begin{aligned}
&K_{\mathrm{HG}}(P_n, P_{\theta^*}; h) \\
&= \frac{1}{n}\sum_{i=1}^n R_i \min(1, h(X_i)) + \frac{1}{n}\sum_{i=1}^n (1-R_i)\min(1, h(X_i)) + \mathrm{E}_{P_{\theta^*}} \min(1, -h(X_i)) \\
&\le \hat{\epsilon} + (1-\hat{\epsilon})\mathrm{E}_{P_{\theta^*,n}} \min(1, h(X_i)) + \mathrm{E}_{P_{\theta^*}} \min(1, -h(X_i)) && (140) \\
&\le \hat{\epsilon} + (1-\hat{\epsilon}) \min(1, \mathrm{E}_{P_{\theta^*,n}} h(X_i)) + \min(1, -\mathrm{E}_{P_{\theta^*}} h(X_i)) && (141) \\
&\le 2\hat{\epsilon} + (1-\hat{\epsilon})\left\{\min(1, \mathrm{E}_{P_{\theta^*,n}} h(X_i)) + \min(1, -\mathrm{E}_{P_{\theta^*}} h(X_i))\right\} && (142) \\
&\le 2\hat{\epsilon} + 0 + |\min(1, -\mathrm{E}_{P_{\theta^*,n}} h(x)) - \min(1, -\mathrm{E}_{P_{\theta^*}} h(x))| && (143) \\
&\le 2\hat{\epsilon} + 0 + |\mathrm{E}_{P_{\theta^*,n}} h(x) - \mathrm{E}_{P_{\theta^*}} h(x)| && (144)
\end{aligned}$$

Lines (140) and (142) hold because $\min(1, u) \vee \min(1, -u) \le 1$ for all $u \in \mathbb{R}$. Line (141) follows from Jensen's inequality by the concavity of $\min(1, u)$ and $\min(1, -u)$. Line (143) follows because $\min(1, u) + \min(1, -u) \le 0$ for all $u \in \mathbb{R}$, and the last line (144) holds because $\min(1, -u)$ is 1-Lipschitz in u. ■

Proposition 52 *In the setting of Proposition 21, it holds with probability at least* $1 - 5\delta$ *that for any* $\gamma \in \Gamma$*,*

$$K_{\mathrm{HG}}(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \le 2(\epsilon + \sqrt{\epsilon/(n\delta)}) + \mathrm{pen}_1(\gamma) C_{\mathrm{sp13}} M_{11} \lambda_{11},$$

where $C_{\mathrm{sp13}} = (5/3)(C_{\mathrm{sp11}} \vee C_{\mathrm{sp12}})$ *with* C_{sp11} *and* C_{sp12} *as in Lemma 30,* $M_{11} = M_1^{1/2}(M_1^{1/2} + 2\sqrt{2\pi})$*, and*

$$\lambda_{11} = \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{n}} + \frac{2\log(5p) + \log(\delta^{-1})}{n}.$$

Proof The proof is similar to that of Proposition 33 and we use the same definition of Ω_1 and Ω_2. In the event Ω_1 we have $|\hat{\epsilon} - \epsilon| \le 1/5$ by the assumption $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \le 1/5$

and hence $\hat{\epsilon} \le 2/5$ by the assumption $\epsilon \le 1/5$. Thus, by Lemma 51 with $\epsilon_1 = 2/5$, it holds in the event Ω_1 that for any $\gamma \in \Gamma$,

$$\begin{aligned} & K_{\mathrm{HG}}(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \\ & \le 2\hat{\epsilon} + \left| \mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\mu^*}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\mu^*}(x) \right| \\ & \le 2(\epsilon + \sqrt{\epsilon/(n\delta)}) + \left| \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x) \right|. \end{aligned} \tag{145}$$

The last step (145) uses the fact that $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\mu^*}(x) = \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x)$ and $\mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\mu^*}(x) = \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*)$, by the definition $h_{\gamma,\mu^*}(x) = h_\gamma(x - \mu^*)$.

Next, as shown in Proposition 33, it holds in the event Ω_2 while conditionally on Ω_1 that for any $\gamma = (\gamma_0, \gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}})^{\mathrm{T}} \in \Gamma$,

$$\begin{aligned} & \left| \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x) \right| \\ & \le \mathrm{pen}_1(\gamma)(5/3)(C_{\mathrm{sp11}} \vee C_{\mathrm{sp12}}) M_{11} \lambda_{11}, \end{aligned} \tag{146}$$

where $h_\gamma(x) = \gamma_0 + \gamma_1^{\mathrm{T}} \varphi(x) + \gamma_2^{\mathrm{T}} (\varphi(x) \otimes \varphi(x))$ and $\mathrm{pen}_1(\gamma) = \|\gamma_1\|_1 + \|\gamma_2\|_1$. Combining (145) and (146) indicates that in the event $\Omega_1 \cap \Omega_2$ with probability at least $1 - 5\delta$, the desired inequality holds for any $\gamma \in \Gamma$. ■

Proposition 53 *In the setting of Proposition 21, it holds with probability at least* $1 - 2\delta$ *that for any* $\gamma \in \Gamma_{\mathrm{rp}}$ *with* $\gamma_0 = 0$ *and* $\mathrm{pen}_1(\gamma) = 1$,

$$\begin{aligned} & K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ & \ge \left\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} - \epsilon - \lambda_{12} \end{aligned}$$

where, with $C_{\mathrm{rad4}} = C_{\mathrm{sg6}} C_{\mathrm{rad3}}$ *is the same constant in Proposition 37,*

$$\lambda_{12} = C_{\mathrm{rad4}} \sqrt{\frac{4 \log(2p(p+1))}{n}} + \sqrt{\frac{2 \log(\delta^{-1})}{n}},$$

depending on the universal constants C_{sg6} *and* C_{rad3} *in Lemma 70 and Corollary 82.*

Proof By definition, for any $\gamma \in \Gamma_{\mathrm{rp}}$, $h_\gamma(x)$ can be represented as $h_{\mathrm{rp},\beta,c}(x)$ such that $\beta_0 = \gamma_0$ and $\mathrm{pen}_1(\beta) = \mathrm{pen}_1(\gamma)$ in the same way as in Proposition 37. Then for any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_1(\gamma) = 1$, we have $\beta_0 = 0$ and $\mathrm{pen}_1(\beta) = 1$ correspondingly, and hence $h_\gamma(x) = h_{\mathrm{rp},\beta,c}(x) \in [-1, 1]$ by the boundedness of the ramp function in $[0, 1]$. Moreover, $h_{\mathrm{rp},\beta,c}(x)$ with $\beta_0 = 0$ and $\mathrm{pen}_1(\beta) = 1$ can be expressed in the form $w^{\mathrm{T}} g(x)$, where for $q = 2p + p(p-1)$, $w \in \mathbb{R}^q$ is an L_1 unit vector, and $g : \mathbb{R}^p \to [0, 1]^q$ is a vector of functions including $\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j - 1)$ for $j = 1, \ldots, p$, and $\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$ for $1 \le i \ne j \le p$. For symmetry, $\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j)\mathrm{ramp}(x_i)$ are included as two distinct components in g, and the corresponding coefficients are identical to each other in w.

Next, $K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ can be bounded as

$$\begin{aligned}
&K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\
&\geq K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \{K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})\}. \qquad (147)
\end{aligned}$$

For any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_1(\gamma) = 1$, because $h_{\gamma,\hat{\mu}}(x) \in [-1, 1]$, we have $\min(h_{\gamma,\hat{\mu}}(x), 1) = h_{\gamma,\hat{\mu}}(x)$ and $\min(-h_{\gamma,\hat{\mu}}(x), 1) = -h_{\gamma,\hat{\mu}}(x)$. Hence the hinge $K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ in (147) reduces to a moment matching term and can be lower bounded as follows:

$$\begin{aligned}
&K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\
&= \mathrm{E}_{P_\epsilon} \min(h_{\gamma,\hat{\mu}}(x), 1) + \mathrm{E}_{P_{\hat{\theta}}} \min(-h_{\gamma,\hat{\mu}}(x), 1) \\
&= \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \\
&= \epsilon \mathrm{E}_Q h_{\gamma,\hat{\mu}}(x) + (1-\epsilon)\mathrm{E}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \\
&\geq -\epsilon + \left\{ \mathrm{E}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\}.
\end{aligned}$$

Similarly, the two other hinge terms in (147) also reduce to moment matching terms:

$$\begin{aligned}
&\{K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})\} \\
&= \left\{ \mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} - \left\{ \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} \\
&= \mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x).
\end{aligned}$$

We apply Lemma 34 with $b = 1$, $g(x) \in [0,1]^q$ defined above, and $f'(\mathrm{e}^u)$ and $f^{\#}(\mathrm{e}^u)$ replaced by the identity function in u. It holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\gamma_0 = 0$ and $\mathrm{pen}_1(\gamma) = b_1$,

$$\begin{aligned}
&\mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x) \\
&\leq C_{\mathrm{sg6}} \sqrt{\frac{V_g \log(2q)}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}} \\
&= C_{\mathrm{sg6}} C_{\mathrm{rad3}} \sqrt{\frac{4\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}},
\end{aligned}$$

as shown in Proposition 37. Combining the preceding three displays leads to the desired result. ■

C.5 Details in main proof of Theorem 16

Proposition 54 *In the setting of Proposition 26, it holds with probability at least* $1 - 4\delta$ *that for any* $\gamma = (\gamma_0, \gamma_1, \gamma_2)^{\mathrm{T}} \in /Gamma,$

$$\begin{aligned}
K_{\mathrm{HG}}(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \leq\ & 2(\epsilon + \sqrt{\epsilon/(n\delta)}) + \mathrm{pen}_2(\gamma_1)(5/3) C_{\mathrm{sp21}} M_2^{1/2} R_1 \lambda_{21} \\
&+ \mathrm{pen}_2(\gamma_2)(25\sqrt{5}/3) C_{\mathrm{sp22}} M_{21} R_1 \sqrt{p} \lambda_{31},
\end{aligned}$$

where C_{sp21} and C_{sp22} are defined as in Lemma 43, $M_{21} = M_2^{1/2}(M_2^{1/2} + 2\sqrt{2\pi})$, and

$$\lambda_{21} = \sqrt{\frac{5p + \log(\delta^{-1})}{n}}, \quad \lambda_{31} = \lambda_{21} + \frac{5p + \log(\delta^{-1})}{n}.$$

Proof The proof is similar to that of Proposition 44 and we use the same definition of Ω_1 and Ω_2. In the event Ω_1 we have $|\hat{\epsilon} - \epsilon| \le 1/5$ by the assumption $\sqrt{\epsilon(1-\epsilon)/(n\delta)} \le 1/5$ and hence $\hat{\epsilon} \le 2/5$ by the assumption $\epsilon \le 1/5$. By Lemma 51 with $\epsilon_1 = 2/5$, it holds that in the event Ω_1 for any $\gamma \in \Gamma$,

$$\begin{aligned}
&K_{\mathrm{HG}}(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \\
&\le 2\hat{\epsilon} + \left| \mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\mu^*}(x) - \mathrm{E}_{P_{\theta^*}} h_{\gamma,\mu^*}(x) \right| \\
&\le 2(\epsilon + \sqrt{\epsilon/(n\delta)}) + \left| \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x) \right|. \qquad (148)
\end{aligned}$$

The last step (148) uses the fact that $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\mu^*}(x) = \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x)$ and $\mathrm{E}_{P_{\theta^*,n}} h_{\gamma,\mu^*}(x) = \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*)$, by the definition $h_{\gamma,\mu^*}(x) = h_\gamma(x - \mu^*)$.

Next, as shown in Proposition 44, it holds in event Ω_2 while conditionally on Ω_1 that for any $\gamma = (\gamma_0, \gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}})^{\mathrm{T}} \in \Gamma$,

$$\begin{aligned}
&\left| \mathrm{E}_{P_{\theta^*,n}} h_\gamma(x - \mu^*) - \mathrm{E}_{P_{(0,\Sigma^*)}} h_\gamma(x) \right| \\
&\le \mathrm{pen}_2(\gamma_1)(5/3) C_{\mathrm{sp21}} M_2^{1/2} \lambda_{21} + \mathrm{pen}_2(\gamma_2)(5/3) C_{\mathrm{sp22}} 3 M_{21} \sqrt{3p} \lambda_{31}, \qquad (149)
\end{aligned}$$

where $h_\gamma(x) = \gamma_0 + \gamma_1^{\mathrm{T}} \varphi(x) + \gamma_2^{\mathrm{T}}(\varphi(x) \otimes \varphi(x))$, $\mathrm{pen}_2(\gamma_1) = \|\gamma_1\|_2$, and $\mathrm{pen}_2(\gamma_2) = \|\gamma_2\|_2$. Combining (148) and (149) indicates that, in the event $\Omega_1 \cap \Omega_2$ with probability at least $1 - 4\delta$, the desired inequality holds for any $\gamma \in \Gamma$. ■

Proposition 55 *In the setting of Proposition 26, it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{10}$,*

$$\begin{aligned}
&K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\
&\ge \left\{ \mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} - \epsilon - \lambda_{22}(2p)^{-1/2}
\end{aligned}$$

where, with $C_{\mathrm{rad5}} = C_{\mathrm{sg},12} C_{\mathrm{rad3}}$, and

$$\lambda_{22} = C_{\mathrm{rad5}} \sqrt{\frac{16p}{n}} + \sqrt{\frac{2p \log(\delta^{-1})}{n}},$$

depending on the universal constants $C_{\mathrm{sg},12}$ and C_{rad3} in Lemma 67 and Corollary 82.

Proof For any $\gamma \in \Gamma_{10} \subset \Gamma_{\mathrm{rp1}}$, because $\mathrm{pen}_2(\gamma) = (2p)^{-1/2}$ and $\Gamma_0 = 0$, we have $\beta_0 = 0$ and $\mathrm{pen}_2(\beta) = p^{-1/2}$. Hence, $h_\gamma(x) = h_{\mathrm{rp1},\beta,c}(x) \in [-1, 1]$ by the Cauchy–Schwartz inequality and the boundedness of the ramp function in $[0, 1]$. Then the mean-centered version $h_{\gamma,\hat{\mu}}(x)$, with $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0$, is also bounded in $[-1, 1]$. Moreover, such $h_{\gamma,\hat{\mu}}(x)$ can be expressed

in the form $\mathrm{pen}_2(\beta) w^{\mathrm{T}}\{g(x-\hat{\mu})-\eta_0\}$, where, for $q=2p$, $w\in\mathbb{R}^q$ is an L_2 unit vector, $g:\mathbb{R}^p\to[0,1]^q$ is a vector of functions, including $\mathrm{ramp}(x_j)$ and $\mathrm{ramp}(x_j-1)$ for $j=1,\ldots,p$, and $\eta_0=\mathrm{E}_{P_{\theta^*}} g(x-\hat{\mu})\in[0,1]^q$.

Next, $K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ can be bounded as

$$\begin{aligned}
&K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\
&\geq K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \{K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})\}. \qquad (150)
\end{aligned}$$

For any $\gamma\in\Gamma_{10}$, because $h_{\gamma,\hat{\mu}}(x)\in[-1,1]$, we have $\min(h_{\gamma,\hat{\mu}}(x),1)=h_{\gamma,\hat{\mu}}(x)$ and $\min(-h_{\gamma,\hat{\mu}}(x),1)=-h_{\gamma,\hat{\mu}}(x)$. Then the hinge term $K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ reduces to a moment matching term and can be lower bounded as follows:

$$\begin{aligned}
&K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\
&= \mathrm{E}_{P_\epsilon} \min(h_{\gamma,\hat{\mu}}(x),1) + \mathrm{E}_{P_{\hat{\theta}}} \min(-h_{\gamma,\hat{\mu}}(x),1) \\
&= \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \\
&= \epsilon \mathrm{E}_Q h_{\gamma,\hat{\mu}}(x) + (1-\epsilon)\mathrm{E}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \\
&\geq -\epsilon + \left\{\mathrm{E}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\right\}. \qquad (151)
\end{aligned}$$

Similarly, the absolute difference term in (150) can be simplified as follows:

$$\begin{aligned}
&\{K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})\} \\
&= \left\{\mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\right\} - \left\{\mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\right\} \\
&= \mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x).
\end{aligned}$$

We apply Lemma 45 with $b=1$, $g(x)\in[0,1]^q$ and η_0 defined above, and $f'(\mathrm{e}^u)$ and $f^{\#}(\mathrm{e}^u)$ replaced by the identity function in u. It holds with probability at least $1-2\delta$ that for any $\gamma\in\Gamma_{01}$,

$$\begin{aligned}
&\mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x) \\
&\leq (2p)^{-1/2}\left\{C_{\mathrm{sg},12}\sqrt{\frac{2qV_g}{n}} + \sqrt{\frac{q\log(\delta^{-1})}{n}}\right\} \\
&\leq (2p)^{-1/2}\left\{C_{\mathrm{sg},12}C_{\mathrm{rad}3}\sqrt{\frac{16p}{n}} + \sqrt{\frac{2p\log(\delta^{-1})}{n}}\right\}, \qquad (152)
\end{aligned}$$

as shown in Proposition 47. Combining the inequalities (150)–(152) leads to the desired result. ■

Proposition 56 *In the setting of Proposition 26, it holds with probability at least* $1-2\delta$ *that for any* $\gamma\in\Gamma_{20}$,

$$\begin{aligned}
&K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\
&\geq \left\{\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x)\right\} - \epsilon - \sqrt{p}\lambda_{32}(4q)^{-1/2}
\end{aligned}$$

where $q = p(1-p)$*, and, with* $C_{\mathrm{rad6}} = C_{\mathrm{sg},12} C_{\mathrm{rad3}}$*,*

$$\lambda_{32} = C_{\mathrm{rad6}} \sqrt{\frac{12(p-1)}{n}} + \sqrt{\frac{(p-1)\log(\delta^{-1})}{n}}.$$

Proof For any $\gamma \in \Gamma_{20} \subset \Gamma_{\mathrm{rp2}}$, because $\mathrm{pen}_2(\gamma) = (4q)^{-1/2}$ and $\gamma_0 = 0$, we have $\mathrm{pen}_2(\beta) = 2\mathrm{pen}_2(\gamma) = q^{-1/2}$. Hence $h_\gamma(x) = h_{\mathrm{rp2},\beta}(x) \in [-1,1]$ by the boundedness of the ramp function in $[0,1]$ and the Cauchy–Schwartz inequality, $\|\beta_2\|_1 \le q^{1/2}\|\beta_2\|_2$. Then the mean-centered version $h_{\gamma,\hat{\mu}}(x)$, with $\mathrm{E}_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) = 0$, is also bounded in $[-1,1]$. Moreover, such $h_{\gamma,\hat{\mu}}(x)$ can be expressed in the form $\mathrm{pen}_2(\beta) w^{\mathrm{T}}\{g(x-\hat{\mu}) - \eta_0\}$, where, for $q = p(p-1)$, $w \in \mathbb{R}^q$ is an L_2 unit vector, $g : \mathbb{R}^p \to [0,1]^q$ is a vector of functions, including $\mathrm{ramp}(x_i)\mathrm{ramp}(x_j)$ for $1 \le i \ne j \le p$, and $\eta_0 = \mathrm{E}_{P_{\theta^*}} g(x-\hat{\mu}) \in [0,1]^q$.

Next, $K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ can be bounded as

$$\begin{aligned} &K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ &\ge K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - |K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})|. \end{aligned} \tag{153}$$

For any $\gamma \in \Gamma_{20}$, because $h_{\gamma,\hat{\mu}}(x) \in [-1,1]$, we have $\min(h_{\gamma,\hat{\mu}}(x), 1) = h_{\gamma,\hat{\mu}}(x)$ and $\min(-h_{\gamma,\hat{\mu}}(x), 1) = -h_{\gamma,\hat{\mu}}(x)$. Then the hinge term $K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})$ reduces to a moment matching term and can be lower bounded as follows:

$$\begin{aligned} &K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \\ &= \mathrm{E}_{P_\epsilon} \min(h_{\gamma,\hat{\mu}}(x), 1) + \mathrm{E}_{P_{\hat{\theta}}} \min(-h_{\gamma,\hat{\mu}}(x), 1) \\ &= \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \\ &= \epsilon \mathrm{E}_Q h_{\gamma,\hat{\mu}}(x) + (1-\epsilon)\mathrm{E}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \\ &\ge -\epsilon + \left\{ \mathrm{E}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\}. \end{aligned} \tag{154}$$

Similarly, the absolute difference term in (153) can be simplified as follows:

$$\begin{aligned} &|K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})| \\ &= \left| \left\{ \mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} - \left\{ \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_{\hat{\theta}}} h_{\gamma,\hat{\mu}}(x) \right\} \right| \\ &= |\mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x)|. \end{aligned}$$

We apply Lemma 45 with $b = 1$, $g(x) \in [0,1]^q$ and η_0 defined above, and $f'(\mathrm{e}^u)$ and $f^{\#}(\mathrm{e}^u)$ replaced by the identity function in u. It holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{20}$,

$$\begin{aligned} &|\mathrm{E}_{P_n} h_{\gamma,\hat{\mu}}(x) - \mathrm{E}_{P_\epsilon} h_{\gamma,\hat{\mu}}(x)| \\ &\le (4q)^{-1/2} \left\{ C_{\mathrm{sg},12} \sqrt{\frac{2qV_g}{n}} + \sqrt{\frac{q\log(\delta^{-1})}{n}} \right\} \\ &= \sqrt{p}(4q)^{-1/2} \left\{ C_{\mathrm{sg},12} C_{\mathrm{rad3}} \sqrt{\frac{12(p-1)}{n}} + \sqrt{\frac{(p-1)\log(\delta^{-1})}{n}} \right\}, \end{aligned} \tag{155}$$

as shown in Proposition 49. Combining the inequalities (153)–(155) leads to the desired result. ■

C.6 Details in proof of Corollary 18

Lemma 57 *Assume that f satisfies Assumptions 1 and 2, and G satisfies Assumption 3. Let $(\hat{\gamma}, \hat{\theta})$ be a solution to the alternating optimization problem (22).*

(i) Let $\epsilon_0 \in (0,1)$ be fixed. For any $\epsilon \in [0, \epsilon_0]$ and any function $h : \mathbb{R}^p \to \mathbb{R}$, we have

$$K_f(P_\epsilon, P_{\hat{\theta}}; h) \le -f'(1-\epsilon_0)\epsilon.$$

(ii) Let $\epsilon_1 \in (0,1)$ be fixed. If $\hat{\epsilon} = n^{-1}\sum_{i=1}^n U_i \in [0, \epsilon_1]$, then for any function $h : \mathbb{R}^p \to \mathbb{R}$, we have

$$K_f(P_n, P_{\hat{\theta}}; h) \le -f'(1-\epsilon_1)\hat{\epsilon} + R_1 |\mathrm{E}_{P_{\theta^*,n}} h(x) - \mathrm{E}_{P_{\theta^*}} h(x)|, \tag{156}$$

where $P_{\theta^*,n}$ denotes the empirical distribution of $\{X_i : U_i = 0, i = 1, \ldots, n\}$ in the latent representation of Huber's contamination model.

Proof (i) As mentioned in Remark 3, the logit f-GAN objective in (10) can be equivalently written as

$$\begin{aligned} K_f(P_\epsilon, P_{\hat{\theta}}; h) &= \mathrm{E}_{P_\epsilon} f'(\mathrm{e}^{h(x)}) - \mathrm{E}_{P_{\hat{\theta}}} f^{\#}(\mathrm{e}^{h(x)}) \\ &= \mathrm{E}_{P_\epsilon} T(h(x)) - \mathrm{E}_{P_{\hat{\theta}}} f^*\{T(h(x))\}, \end{aligned}$$

where $T(u) = f'(\mathrm{e}^u)$ and f^* is the convex conjugate of f. Because f is convex and non-decreasing by Assumptions 1 and 2, we have that f^* is convex and non-decreasing.

Denote as $L_G(\theta, \gamma)$ the generator objective function, $\mathrm{E}_{P_\epsilon} f'(\mathrm{e}^{h_{\gamma,\mu}(x)}) - \mathrm{E}_{P_\theta} G(h_{\gamma,\mu}(x))$. By the definition of a solution to alternating optimization (see Remark 1), we have

$$\begin{aligned} L_G\{(\hat{\mu}, \hat{\Sigma}), \hat{\gamma}\} &\le L_G\{(\mu^*, \hat{\Sigma}), \hat{\gamma}\}, \\ L_G\{(\hat{\mu}, \hat{\Sigma}), \hat{\gamma}\} &\le L_G\{(\hat{\mu}, \Sigma^*), \hat{\gamma}\}, \end{aligned}$$

that is, the generator loss at $\hat{\theta}$ is less than that at any θ, with the discriminator parameter fixed at $\hat{\gamma}$. The preceding inequalities can be written out as

$$\mathrm{E}_{P_\epsilon} T(\mathrm{e}^{h_{\hat{\gamma},\hat{\mu}}(x)}) - \mathrm{E}_{P_{\hat{\theta}}} G(h_{\hat{\gamma},\hat{\mu}}(x)) \le \mathrm{E}_{P_\epsilon} T(\mathrm{e}^{h_{\hat{\gamma},\mu^*}(x)}) - \mathrm{E}_{P_{\mu^*,\hat{\Sigma}}} G(h_{\hat{\gamma},\mu^*}(x)),$$

and

$$\mathrm{E}_{P_\epsilon} T(\mathrm{e}^{h_{\hat{\gamma},\hat{\mu}}(x)}) - \mathrm{E}_{P_{\hat{\theta}}} G(h_{\hat{\gamma},\hat{\mu}}(x)) \le \mathrm{E}_{P_\epsilon} T(\mathrm{e}^{h_{\hat{\gamma},\hat{\mu}}(x)}) - \mathrm{E}_{P_{\hat{\mu},\Sigma^*}} G(h_{\hat{\gamma},\hat{\mu}}(x)).$$

Note that either the location or the variance matrix, but not both, is changed on the two sides in each inequality. In the first inequality, we have $\mathrm{E}_{P_{\hat{\theta}}} G(h_{\hat{\gamma},\hat{\mu}}(x)) = \mathrm{E}_{P_{\mu^*,\hat{\Sigma}}} G(h_{\hat{\gamma},\mu^*}(x))$ because both are equal to $\mathrm{E}_{P_{0,\hat{\Sigma}}} G(h_{\hat{\gamma},0}(x))$. In the second inequality, the term $\mathrm{E}_{P_\epsilon} T(\mathrm{e}^{h_{\hat{\gamma},\hat{\mu}}(x)})$ is on both sides. Then the two inequalities yield

$$\begin{aligned} \mathrm{E}_{P_\epsilon} T(\mathrm{e}^{h_{\hat{\gamma},\hat{\mu}}(x)}) &\le \mathrm{E}_{P_\epsilon} T(\mathrm{e}^{h_{\hat{\gamma},\mu^*}(x)}), \\ -\mathrm{E}_{P_{\hat{\theta}}} G(h_{\hat{\gamma},\hat{\mu}}(x)) &\le -\mathrm{E}_{P_{\hat{\mu},\Sigma^*}} G(h_{\hat{\gamma},\hat{\mu}}(x)). \end{aligned}$$

Now we are ready to derive an upper bound for $K_f(P_\epsilon, P_{\hat\theta}; h_{\hat\gamma,\hat\mu})$:

$$\begin{aligned}
&K_f(P_\epsilon, P_{\hat\theta}; h_{\hat\gamma,\hat\mu}) \\
&= \mathrm{E}_{P_\epsilon} T(h_{\hat\gamma,\hat\mu}) - \mathrm{E}_{P_{\hat\theta}} f^*\{T(G^{-1}(G(h_{\hat\gamma,\hat\mu})))\} \\
&\le \mathrm{E}_{P_\epsilon} T(h_{\hat\gamma,\hat\mu}) - f^*\{T(G^{-1}(\mathrm{E}_{P_{\hat\theta}} G(h_{\hat\gamma,\hat\mu})))\} && (157)\\
&\le \mathrm{E}_{P_\epsilon} T(h_{\hat\gamma,\mu^*}) - f^*\{T(G^{-1}(\mathrm{E}_{P_{\hat\mu,\Sigma^*}} G(h_{\hat\gamma,\hat\mu})))\} && (158)\\
&\le (1-\epsilon)\mathrm{E}_{P_{\theta^*}} T(h_{\hat\gamma,\mu^*}) - f^*\{T(G^{-1}(\mathrm{E}_{P_{\mu^*,\Sigma^*}} G(h_{\hat\gamma,\mu^*})))\} && (159)\\
&\le (1-\epsilon) T(\mathrm{E}_{P_{\theta^*}} h_{\hat\gamma,\mu^*}) - f^*\{T(\mathrm{E}_{P_{\mu^*,\Sigma^*}} h_{\hat\gamma,\mu^*})\} && (160)\\
&\le f(1-\epsilon) \le -f'(1-\epsilon_0)\epsilon. && (161)
\end{aligned}$$

Line (157) follows from Jensen's inequality by the convexity of f^*. Line (158) follows from the two inequalities derived above, together with the fact that $-f^*(T(G^{-1}))$ is non-increasing, by non-decreasingness of f^*, T, and G^{-1}. In (159) we use the fact that $\mathrm{E}_{P_{\hat\mu,\Sigma^*}} G(h_{\hat\gamma,\hat\mu}) = \mathrm{E}_{P_{\mu^*,\Sigma^*}} G(h_{\hat\gamma,\mu^*}(x))$ and drop the E_Q term because $T \le 0$. Line (160) follows from Jensen's inequality by the convexity of G and the concavity of T, together with the fact that $-f^*(T(G^{-1}))$ is non-increasing. For the last line (161), by the definition of Fenchel conjugate we have

$$(1-\epsilon)s - f^*(s) \le f(1-\epsilon) \le -f'(1-\epsilon_0)\epsilon,$$

with s set to $\mathrm{E}_{P_{\theta^*}} T(h_{\hat\gamma,\mu^*})$.

(ii) To derive an upper bound for $K_f(P_n, P_{\hat\theta}; h_{\hat\gamma})$, we first argue similarly as in part (i):

$$\begin{aligned}
&K_f(P_n, P_{\hat\theta}; h_{\hat\gamma,\hat\mu}) \\
&= \hat\epsilon\, \mathrm{E}_Q T(h_{\hat\gamma,\hat\mu}) + (1-\hat\epsilon)\mathrm{E}_{P_{\theta^*,n}} T(h_{\hat\gamma,\hat\mu}) - \mathrm{E}_{P_{\hat\theta}} f^*\{T(h_{\hat\gamma,\hat\mu})\} \\
&\le (1-\hat\epsilon) T(\mathrm{E}_{P_{\theta^*,n}} h_{\hat\gamma,\mu^*}) - f^*\{T(\mathrm{E}_{P_{\mu^*,\Sigma^*}} h_{\hat\gamma,\mu^*})\}.
\end{aligned}$$

Then we use the R_1-Lipschitz property of $f^*(T)$ and obtain

$$\begin{aligned}
&(1-\hat\epsilon) T(\mathrm{E}_{P_{\theta^*,n}} h_{\hat\gamma,\mu^*}) - f^*\{T(\mathrm{E}_{P_{\mu^*,\Sigma^*}} h_{\hat\gamma,\mu^*})\} \\
&\le (1-\hat\epsilon) T(\mathrm{E}_{P_{\theta^*,n}} h_{\hat\gamma,\mu^*}) - f^*\{T(\mathrm{E}_{P_{\theta^*,n}} h_{\hat\gamma,\mu^*})\} + R_1 |\mathrm{E}_{P_{\theta^*,n}} h_{\hat\gamma,\mu^*} - \mathrm{E}_{P_{\theta^*}} h_{\hat\gamma,\mu^*}| \\
&\le f(1-\hat\epsilon) + R_1 |\mathrm{E}_{P_{\theta^*,n}} h_{\hat\gamma,\mu^*} - \mathrm{E}_{P_{\theta^*}} h_{\hat\gamma,\mu^*}| \\
&\le -f'(1-\epsilon_1)\hat\epsilon + R_1 |\mathrm{E}_{P_{\theta^*,n}} h_{\hat\gamma,\mu^*} - \mathrm{E}_{P_{\theta^*}} h_{\hat\gamma,\mu^*}|.
\end{aligned}$$

Combining the preceding displays completes the proof. ■

Lemma 58 *Let $(\hat\gamma, \hat\theta)$ be a solution to the alternating optimization problem (23).*

(i) For any $\epsilon \in [0,1]$ and any function $h : \mathbb{R}^p \to \mathbb{R}$, we have

$$K_{\mathrm{HG}}(P_\epsilon, P_{\hat\theta}; h) \le 2\epsilon.$$

(ii) If $\hat\epsilon = n^{-1}\sum_{i=1}^n U_i \in [0,1]$, then for any function $h : \mathbb{R}^p \to \mathbb{R}$, we have

$$K_{\mathrm{HG}}(P_n, P_{\hat\theta}; h) \le 2\hat\epsilon + |\mathrm{E}_{P_{\theta^*,n}} h(x) - \mathrm{E}_{P_{\theta^*}} h(x)|, \tag{162}$$

where $P_{\theta^*,n}$ denotes the empirical distribution of $\{X_i : U_i = 0, i = 1, \ldots, n\}$ in the latent representation of Huber's contamination model.

Proof (i) By the same argument used in the proof of Lemma 57 with $T(u)$ replaced by $\min(u,1)$ and $-f^*(T(u))$ replaced by $\min(-u,1)$ we have:

$$\begin{aligned}
&K_{\mathrm{HG}}(P_\epsilon, P_{\hat{\theta}}; h_{\hat{\gamma},\hat{\mu}}) \\
&\leq \epsilon + (1-\epsilon)\min(\mathrm{E}_{P_{\theta^*}} h_{\hat{\gamma},\mu^*}, 1) + \min(-\mathrm{E}_{P_{\theta^*}} h_{\hat{\gamma},\mu^*}, 1) \qquad (163)\\
&\leq 2\epsilon. \qquad (164)
\end{aligned}$$

Inequality (163) is derived by the same argument that leads to (157)-(160) with the fact that $\min(u,1)$ is concave and non-decreasing and that $\min(-u,1)$ is concave and non-increasing just like T and $-f^*(T)$ in Lemma 57 respectively. The ϵ term is a result of the fact that $\min(u,1) \leq 1$. Inequality (164) is by the same argument used in Lemma 51:

$$\begin{aligned}
&(1-\epsilon)\min(u,1) + \min(-u,1) \\
&\leq \epsilon + (1-\epsilon)\{\min(u,1) + \min(-u,1)\} \\
&\leq \epsilon,
\end{aligned}$$

with u set to be $h_{\hat{\gamma},\mu^*}$.

(ii) To derive an upper bound for $K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\hat{\gamma}})$, we first argue similarly as in part (i):

$$\begin{aligned}
&K_{\mathrm{HG}}(P_n, P_{\hat{\theta}}; h_{\hat{\gamma},\hat{\mu}}) \\
&= \hat{\epsilon}\,\mathrm{E}_Q \min(h_{\hat{\gamma},\hat{\mu}}, 1) + (1-\hat{\epsilon})\mathrm{E}_{P_{\theta^*,n}} \min(h_{\hat{\gamma},\hat{\mu}}, 1) + \mathrm{E}_{P_{\hat{\theta}}} \min(-h_{\hat{\gamma},\hat{\mu}}, 1) \\
&\leq \hat{\epsilon} + (1-\hat{\epsilon})\min(\mathrm{E}_{P_{\theta^*,n}} h_{\hat{\gamma},\mu^*}, 1) + \min(-\mathrm{E}_{P_{\mu^*,\Sigma^*}} h_{\hat{\gamma},\mu^*}, 1).
\end{aligned}$$

Then we use the 1-Lipschitz property of $\min(-u,1)$ and obtain

$$\begin{aligned}
&(1-\hat{\epsilon})\min(\mathrm{E}_{P_{\theta^*,n}} h_{\hat{\gamma},\mu^*}, 1) + \min(-\mathrm{E}_{P_{\mu^*,\Sigma^*}} h_{\hat{\gamma},\mu^*}, 1) \\
&\leq (1-\hat{\epsilon})\min(\mathrm{E}_{P_{\theta^*,n}} h_{\hat{\gamma},\mu^*}, 1) + \min(-\mathrm{E}_{P_{\theta^*,n}} h_{\hat{\gamma},\mu^*}, 1) + |\mathrm{E}_{P_{\theta^*,n}} h_{\hat{\gamma},\mu^*} - \mathrm{E}_{P_{\theta^*}} h_{\hat{\gamma},\mu^*}| \\
&\leq \hat{\epsilon} + |\mathrm{E}_{P_{\theta^*,n}} h_{\hat{\gamma},\mu^*} - \mathrm{E}_{P_{\theta^*}} h_{\hat{\gamma},\mu^*}|.
\end{aligned}$$

Combining the preceding displays completes the proof. ■

C.7 Proofs in Section 5

Proof [Proof of Proposition 19] We first verify that $f(t) = \frac{1+t}{2} g_0(\frac{2t}{1+t})$ is convex on $[0,+\infty)$ and $f(1) = 0$ so that D_f is a valid f-divergence. Because g_0 is convex by the convexity of g and $f''(t) = \frac{2}{(1+t)^3} g_0''(\frac{2t}{1+t})$, it follows that f is convex on $[0,+\infty)$. Direct calculation gives $f(1) = g_0(1) = 0$. Thus, f defines a valid f-divergence.

Next we show that $L_g(P_*, P_\theta; q_\gamma) = K_f(P_*, P_\theta; h_\gamma)$. Denote $\mathrm{e}^{h_\gamma(x)}$ by t and $q_\gamma(x)$ by q. Then we have

$$\begin{aligned}
&K_f(P_*, P_\theta; h_\gamma) = \mathrm{E}_{P_*} f'(t) - \mathrm{E}_{P_\theta}\left\{tf'(t) - f(t)\right\} \\
&= \mathrm{E}_{P_*}\left\{\frac{1}{1+t} g_0'\left(\frac{2t}{1+t}\right) + \frac{1}{2} g_0\left(\frac{2t}{1+t}\right)\right\} - \mathrm{E}_{P_\theta}\left\{\frac{t}{1+t} g_0'\left(\frac{2t}{1+t}\right) - \frac{1}{2} g_0\left(\frac{2t}{1+t}\right)\right\} \qquad (165) \\
&= \mathrm{E}_{P_*}\left\{(1-q) g_0'(2q) + \frac{1}{2} g_0(2q)\right\} - \mathrm{E}_{P_\theta}\left\{q g_0'(2q) - \frac{1}{2} g_0(2q)\right\} \\
&= \mathrm{E}_{P_*}\left\{\frac{1-q}{2} g'(q) + \frac{1}{2} g(q) - \frac{1}{2} g\left(\frac{1}{2}\right)\right\} - \mathrm{E}_{P_\theta}\left\{\frac{q}{2} g'(q) - \frac{1}{2} g(q) + \frac{1}{2} g\left(\frac{1}{2}\right)\right\} \qquad (166) \\
&= \frac{1}{2}\left\{\mathrm{E}_{P_*} S_g(q, 1) - \mathrm{E}_{P_\theta} S_g(q, 0)\right\} - g\left(\frac{1}{2}\right). \qquad (167)
\end{aligned}$$

Line (165) is by direct calculation. Lines (166)–(167) are by the definition of g and S_g.

Finally, by the definition of f from g_0, direct calculation gives

$$\int q f\left(\frac{p}{q}\right) = \int \frac{p+q}{2} g_0\left(\frac{2p}{p+q}\right),$$

which implies that $D_{g_0}(P_* || (P_* + P_\theta)/2) = D_f(P_* || P_\theta)$. ■

Appendix D. Auxiliary lemmas

D.1 Truncated linear basis

The following result gives upper bounds on the moments of the truncated linear basis, which are used in the proofs of Lemma 30 and 43.

Lemma 59 *For $X \sim \mathrm{N}(0, \sigma^2)$ and $\xi \in \mathbb{R}$, we have*

$$\begin{aligned}
&\mathrm{E}(X - \xi)_+ \le \frac{\sigma}{\sqrt{2\pi}} + |\xi|, \\
&\mathrm{E}[\{(X - \xi)_+\}^2] \le \sigma^2 + \xi^2.
\end{aligned}$$

Proof The second result is immediate: $\mathrm{E}[\{(X-\xi)_+\}^2] \le \mathrm{E}\{(X-\xi)^2\} = \sigma^2 + \xi^2$. The first result can be shown as follows:

$$\begin{aligned}
&\mathrm{E}(X - \xi)_+ = \int_\xi^\infty (x - \xi) \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x \\
&= \frac{\sigma}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\xi^2}{2\sigma^2}} - \xi \int_\xi^\infty \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x \\
&\le \frac{\sigma}{\sqrt{2\pi}} + |\xi|.
\end{aligned}$$

The last inequality holds because $0 \le \int_\xi^\infty \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x \le 1$. ■

D.2 VC index of ramp functions

For a collection $\mathcal{C}$ of subsets of $\mathcal{X}$, and points $x_1, \ldots, x_n \in \mathcal{X}$, define

$$\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) = \#\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\},$$

that is, $\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)$ is the number of subsets of $\{x_1, \ldots, x_n\}$ picked out by the collection $\mathcal{C}$. We say that a subset $\{x_i, \ldots, x_j\} \subset \{x_1, \ldots, x_n\}$ is picked up by $\mathcal{C}$ if $\{x_i, \ldots, x_j\} \in \{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}$. For convenience, we also say that $\{x_i, \ldots, x_j\}$ is picked up by C if $\{x_i, \ldots, x_j\} = C \cap \{x_1, \ldots, x_n\}$. Moreover, define

$$m^{\mathcal{C}}(n) = \max_{x_1, \ldots, x_n} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n),$$

and the Vapnik–Chervonenkis (VC) index of $\mathcal{C}$ as (van der Vaart and Wellner, 1996)

$$V(\mathcal{C}) = \inf\{n \geq 1 : m^{\mathcal{C}}(n) < 2^n\}.$$

where the infimum over the empty set is taken to be infinity.

The subgraph of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as $G_f = \{(x, t) \in \mathcal{X} \times \mathbb{R} : t < f(x)\}$. For a collection of functions $\mathcal{F}$, denote the collection of corresponding subgraphs as $G_{\mathcal{F}} = \{G_f : f \in \mathcal{F}\}$, and define the VC index of $\mathcal{F}$ as $V(\mathcal{F}) = V(G_{\mathcal{F}})$.

Lemma 60 *For $\mathcal{F} = \{f_b(x) = 1 - (x + b)_+ + (x + b - 1)_+ : b \in \mathbb{R}\}$, we have that $V(\mathcal{F}) = 2$. Moreover, the VC index of $\{f_b(-x) : b \in \mathbb{R}\} = \{\text{ramp}(x - b) : b \in \mathbb{R}\}$ is 2. That is, the VC index of moving-knots ramp functions is 2.*

Proof

To show $V(\mathcal{F}) = 2$, we need to show $m^{G_{\mathcal{F}}}(1) = 2^1$ and $m^{G_{\mathcal{F}}}(2) < 2^2$. The first property is trivially true. For the second property, it suffices to show that for any two distinct points $\{(x_1, t_1), (x_2, t_2)\}$, there is at least a subset of $\{(x_1, t_1), (x_2, t_2)\}$ that cannot be picked up by G_{f_b} for any $f_b \in \mathcal{F}$. Without loss of generality, assume that $x_1 \leq x_2$. We arbitrarily fix $f_b \in \mathcal{F}$ and discuss several cases depending on $t_2 - t_1$ and $x_2 - x_1$.

If $t_2 - t_1 \leq -(x_2 - x_1)$, then if $(x_1, t_1) \in G_{f_b}$, i.e., $f_b(x_1) > t_1$, we have

$$\begin{aligned} f_b(x_2) - f_b(x_1) &\geq (-1)(x_2 - x_1) \\ &\geq t_2 - t_1. \end{aligned}$$

This implies that $f_b(x_2) \geq f_b(x_1) - t_1 + t_2 > t_2$ and hence $(x_2, t_2) \in G_{f_b}$. As a result, a subset containing just (x_1, t_1) cannot be picked up by G_{f_b}.

If $t_2 - t_1 \geq 0$, then if $(x_2, t_2) \in G_{f_b}$, i.e., $f_b(x_2) > t_2$, we have

$$f_b(x_1) \geq f_b(x_2) > t_2 \geq t_1,$$

and hence $(x_2, t_2) \in G_{f_b}$. Thus, the subset $\{(x_2, t_2)\}$ can never be picked up by G_{f_b}.

If $t_2 - t_1 < 0$ and $t_2 - t_1 > -(x_2 - x_1)$, then if $f_b(x_2) > t_2$, we have

$$\begin{aligned} f_b(x_1) - f_b(x_2) &\geq (-1)(x_1 - x_2) \\ &> -(t_2 - t_1). \end{aligned}$$

This implies that $f_b(x_1) > f_b(x_2) - t_2 + t_1 > t_1$. As a result, the subset $\{(x_2, t_2)\}$ can never be picked up by G_{f_b}.

Combining the preceding cases shows that $m^{G_{\mathcal{F}}}(2) < 2^2$ and $V(\mathcal{F}) = 2$. Moreover, the class of functions $\{f_b(-x) : b \in \mathbb{R}\}$, denoted as $\tilde{\mathcal{F}}$, admits a one-to-one correspondence with $\mathcal{F}$. A subset of $\{(x_1, t_1), (x_2, t_2)\}$ is picked up by $G_{\mathcal{F}}$ if and only if the corresponding subset of $\{(-x_1, t_1), (-x_2, t_2)\}$ is picked up by $G_{\tilde{\mathcal{F}}}$. Hence $V(\tilde{\mathcal{F}}) = V(\mathcal{F}) = 2$. ■

Lemma 61 *For $\mathcal{F} = \{f(x) \equiv b : b \in \mathbb{R}\}$, we have that $V(\mathcal{F}) = 2$. That is, the VC index of constant functions is 2.*

Proof For any two distinct points (x_1, t_1) and (x_2, t_2), assume that with loss of generality $t_1 \le t_2$. Then the singleton $\{(x_2, t_2)\}$ can never be picked up by G_f for any $f \in \mathcal{F}$, and hence $m^{G_{\mathcal{F}}}(2) < 2^2$ and $V(\mathcal{F}) = 2$. In fact, if (x_2, t_2) is in the subgraph of $f(x) \equiv b$, then $t_2 < b$. As a result, $t_1 \le t_2 < b$, indicating that (x_1, t_1) is also in the subgraph. ■

D.3 Lipschitz functions of Gaussian vectors

Say that a function $g : \mathbb{R}^p \to \mathbb{R}^m$ is L-Lipschitz if $\|g(x_1) - g(x_2)\|_2 \le L\|x_1 - x_2\|_2$ for any $x_1, x_2 \in \mathbb{R}^p$.

Lemma 62 *Let $X \sim \mathrm{N}_p(\mu, \Sigma)$, and $g : \mathbb{R}^p \to \mathbb{R}^m$ be an L-Lipschitz function.*

(i) For any vector $w \in \mathbb{R}^m$ with $\|w\|_2 = 1$, we have

$$\mathrm{E}\left[\{w^{\mathrm{T}}(g(X) - \mathrm{E}g(X))\}^2\right] \le 2C_{\mathrm{sg},12}^2 L^2 \|\Sigma\|_{\mathrm{op}},$$

where $C_{\mathrm{sg},12}$ is the universal constant from Lemma 67. Hence we have

$$\|\mathrm{Var}\, g(X)\|_{\mathrm{op}} \le 2C_{\mathrm{sg},12}^2 L^2 \|\Sigma\|_{\mathrm{op}}.$$

(ii) For any symmetric matrix $A \in \mathbb{R}^{m \times m}$ with $\|A\|_{\mathrm{F}} = 1$, we have

$$\mathrm{E}\left[\{(g(X) - \mathrm{E}g(X))^{\mathrm{T}} A (g(X) - \mathrm{E}g(X))\}^2\right] \le 4C_{\mathrm{sg},12}^4 m L^4 \|\Sigma\|_{\mathrm{op}}^2.$$

Proof (i) By Boucheron et al. (2013), Theorem 5.6, it can be shown that for any L_2 unit vector w, $w^{\mathrm{T}}(g(X) - \mathrm{E}g(X))$ is sub-gaussian with tail parameter $L\|\Sigma\|_{\mathrm{op}}^{1/2}$. See the proof of Lemma 43(i) for a similar argument. Then by Lemma 67, $\mathrm{E}[\{w^{\mathrm{T}}(g(X) - \mathrm{E}g(X))\}^2] \le 2C_{\mathrm{sg},12}^2 L^2 \|\Sigma\|_{op}$.

(ii) Consider an eigen-decomposition $A = \sum_{j=1}^m \lambda_j w_j w_j^{\mathrm{T}}$, where λ_j's are eigenvalues and w_j's are the eigenvectors with $\|w_j\|_2 = 1$. Denote $g = g(X)$ and $\tilde{g} = g - \mathrm{E}g$. Then $\tilde{g}^{\mathrm{T}} A \tilde{g} = \sum_{j=1}^m \lambda_j (w_j^{\mathrm{T}} \tilde{g})^2$ and

$$(\tilde{g}^{\mathrm{T}} A \tilde{g})^2 \le \left(\sum_{j=1}^m \lambda_j^2\right)\left(\sum_{j=1}^m (w_j^{\mathrm{T}} \tilde{g})^4\right) \le 4C_{\mathrm{sg},12}^4 m L^4 \|\Sigma\|_{\mathrm{op}}^2.$$

The first step uses the Cauchy–Schwartz inequality, and the second step uses the fact that $\sum_{j=1}^m \lambda_j^2 = \|A\|_{\mathrm{F}}^2 = 1$ and $E(w_j^{\mathrm{T}}\tilde{g})^4 \le 4C_{\mathrm{sg},12}^4 L^4 \|\Sigma\|_{\mathrm{op}}^2$ by Lemma 67 because $w_j^{\mathrm{T}}\tilde{g}_j$ is sub-gaussian with tail parameter $L\|\Sigma\|_{\mathrm{op}}^{1/2}$ for each j. ■

The following result provides a 4th-moment bound which depends linearly on $L^2\|\Sigma\|_{\mathrm{op}}$, under a boundedness condition in addition to the Lipschitz condition.

Lemma 63 *Let $X \sim \mathrm{N}_p(\mu, \Sigma)$, and $g : \mathbb{R}^p \to [0,1]^m$ be an L-Lipschitz function.*
(i) For any matrix $A \in \mathbb{R}^{m\times m}$ with $\|A\|_{\mathrm{F}} = 1$, we have

$$\mathrm{E}\left[\{(g(X) - \mathrm{E}g(X))^{\mathrm{T}} A (g(X) - \mathrm{E}g(X))\}^2\right] \le 2C_{\mathrm{sg},12}^2 m L^2 \|\Sigma\|_{\mathrm{op}}.$$

(ii) For any matrix $A \in \mathbb{R}^{m\times m}$ with $\|A\|_{\mathrm{F}} = 1$, we have

$$\mathrm{E}\left[\{g^{\mathrm{T}}(X)Ag(X) - \mathrm{E}g^{\mathrm{T}}(X)Ag(X)\}^2\right] \le 20C_{\mathrm{sg},12}^2 m L^2 \|\Sigma\|_{\mathrm{op}}.$$

Proof (i) Denote $g = g(X)$ and $\tilde{g} = g - \mathrm{E}g$. Then each component of $\tilde{g}$ is contained in $[-1,1]$ by the boundedness of g. The variable $(\tilde{g}^{\mathrm{T}}A\tilde{g})^2$ can be bounded as follows:

$$\begin{aligned}
&(\tilde{g}^{\mathrm{T}}A\tilde{g})^2 = \mathrm{tr}(\tilde{g}^{\mathrm{T}}A\tilde{g}\tilde{g}^{\mathrm{T}}A^{\mathrm{T}}\tilde{g}) = \mathrm{tr}(A\tilde{g}\tilde{g}^{\mathrm{T}}A^{\mathrm{T}}\tilde{g}\tilde{g}^{\mathrm{T}}) \\
&\le \mathrm{tr}(A\tilde{g}\tilde{g}^{\mathrm{T}}A^{\mathrm{T}})\|\tilde{g}\tilde{g}^{\mathrm{T}}\|_{\mathrm{op}} \qquad (168)\\
&\le m\,\mathrm{tr}(A\tilde{g}\tilde{g}^{\mathrm{T}}A^{\mathrm{T}}) = m\,\mathrm{tr}(A^{\mathrm{T}}A\tilde{g}\tilde{g}^{\mathrm{T}}). \qquad (169)
\end{aligned}$$

Line (168) follows from von Neumann's trace equality. Line (169) uses the fact that $\|\tilde{g}\tilde{g}^{\mathrm{T}}\|_{\mathrm{op}} \le m$, because $w^{\mathrm{T}}\tilde{g}\tilde{g}^{\mathrm{T}}u = (w^{\mathrm{T}}\tilde{g})^2 \le \|w\|_2^2\|\tilde{g}\|_2^2 \le m\|w\|_2^2$ for any $w \in \mathbb{R}^m$, by the boundedness of $\tilde{g}$. Then the desired result follows because

$$\begin{aligned}
&\mathrm{E}\mathrm{tr}(A^{\mathrm{T}}A\tilde{g}\tilde{g}^{\mathrm{T}}) = \mathrm{tr}(A^{\mathrm{T}}A\,\mathrm{Var}(g)) \\
&\le \mathrm{tr}(A^{\mathrm{T}}A)\|\mathrm{Var}(g)\|_{\mathrm{op}} \qquad (170)\\
&\le 2C_{\mathrm{sg},12}^2 L^2 \|\Sigma\|_{\mathrm{op}}. \qquad (171)
\end{aligned}$$

Line (170) also follows from von Neumann's trace equality. Line (171) follows because $\mathrm{tr}(A^{\mathrm{T}}A) = \|A\|_{\mathrm{F}}^2 = 1$ and $\|\mathrm{Var}(g)\|_{\mathrm{op}} \le 2C_{\mathrm{sg},12}^2 L^2\|\Sigma\|_{\mathrm{op}}$ by Lemma 62(i), with g being an L-Lipschitz function.

(ii) The difference $g^{\mathrm{T}}Ag - \mathrm{E}g^{\mathrm{T}}Ag$ can be expressed in terms of the centered variables as $g^{\mathrm{T}}Ag - \mathrm{E}g^{\mathrm{T}}Ag = (\tilde{g}^{\mathrm{T}}A\tilde{g} - \mathrm{E}\tilde{g}^{\mathrm{T}}A\tilde{g}) + 2(\mathrm{E}g)^{\mathrm{T}}A\tilde{g}$. Then

$$\begin{aligned}
&(g^{\mathrm{T}}Ag - \mathrm{E}g^{\mathrm{T}}Ag)^2 \le 2(\tilde{g}^{\mathrm{T}}A\tilde{g} - \mathrm{E}\tilde{g}^{\mathrm{T}}A\tilde{g})^2 + 8\{(\mathrm{E}g)^{\mathrm{T}}A\tilde{g}\}^2 \\
&\le 2(\tilde{g}^{\mathrm{T}}A\tilde{g} - \mathrm{E}\tilde{g}^{\mathrm{T}}A\tilde{g})^2 + 8\|\mathrm{E}g\|_2^2\|A\tilde{g}\|_2^2. \qquad (172)
\end{aligned}$$

The expectation of the first term on (172) can be bounded using (i) as

$$\begin{aligned}
&2\mathrm{E}\{(\tilde{g}^{\mathrm{T}}A\tilde{g} - \mathrm{E}\tilde{g}^{\mathrm{T}}A\tilde{g})^2\} = 2\mathrm{E}\{(\tilde{g}^{\mathrm{T}}A\tilde{g})^2\} - 2(\mathrm{E}\tilde{g}^{\mathrm{T}}A\tilde{g})^2 \\
&\le 2\mathrm{E}\{(\tilde{g}^{\mathrm{T}}A\tilde{g})^2\} \le 4mC_{\mathrm{sg},12}^2 L^2\|\Sigma\|_{\mathrm{op}}.
\end{aligned}$$

The expectation of the second term on (172) can be bounded as

$$\begin{aligned} &8\|\mathrm{E}g\|_2^2\,\mathrm{E}\|A\tilde{g}\|_2^2 = 8\|\mathrm{E}g\|_2^2\,\mathrm{Etr}(A^{\mathrm{T}}A\tilde{g}\tilde{g}^{\mathrm{T}}) \\ &\leq 16mC_{\mathrm{sg},12}^2L^2\|\Sigma\|_{\mathrm{op}}. \end{aligned}$$

by inequality (171) and the fact that $\|\mathrm{E}g\|_2^2 \leq m$. Combining the preceding two bounds yields the desired result. ■

D.4 Moment matching for Lipschitz functions

The following result gives an upper bound on moment matching of quadratic forms under a Lipschitz condition.

Lemma 64 *Let $g : \mathbb{R}^p \to \mathbb{R}^m$ be an L-Lipschitz function, and let $X_1 = \mu_1 + D_1Z$ and $X_2 = \mu_2 + D_2Z$, where $Z \in \mathbb{R}^p$ is a random vector in which the second moments of all components are 1, $\mu_1, \mu_2 \in \mathbb{R}^p$, and $D_1 = \mathrm{diag}(d_1)$ and $D_2 = \mathrm{diag}(d_2)$ with $d_1, d_2 \in \mathbb{R}_+^p$. Then for any matrix $A \in \mathbb{R}^{m\times m}$ with $\|A\|_{\mathrm{F}} = 1$,*

$$\begin{aligned} &|\mathrm{E}g^{\mathrm{T}}(X_1)Ag(X_1) - \mathrm{E}g^{\mathrm{T}}(X_2)Ag(X_2)| \\ &\leq 2\sqrt{m}L^2\Delta^2 + 2\sqrt{2}L\Delta(\mathrm{E}\|g_2\|_2^2)^{1/2}, \end{aligned}$$

where $\Delta^2 = \|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2$.

Proof Denote $g_1 = g(X_1)$ and $g_2 = g(X_2)$. The difference $g_1^{\mathrm{T}}Ag_1 - g_2^{\mathrm{T}}Ag_2$ can be decomposed as

$$\begin{aligned} &g_1^{\mathrm{T}}Ag_1 - g_2^{\mathrm{T}}Ag_2 \\ &= (g_1 - g_2)^{\mathrm{T}}A(g_1 - g_2) + 2(g_1 - g_2)^{\mathrm{T}}Ag_2. \end{aligned} \tag{173}$$

The expectation of the first term on (173) can be bounded as

$$\begin{aligned} &|\mathrm{E}(g_1 - g_2)^{\mathrm{T}}A(g_1 - g_2)| = |\mathrm{tr}\,(AV)| \\ &\leq \sum_{j=1}^m s_j(A)\|V\|_{\mathrm{op}} &(174)\\ &\leq 2\sqrt{m}L^2\left(\|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2\right), &(175) \end{aligned}$$

where $s_1(A), \ldots, s_m(A)$ are the singular values of A, and $V = \mathrm{E}\{(g_1 - g_2)(g_1 - g_2)^{\mathrm{T}}\}$. Line (174) follows from von Neumann's trace inequality. Line (175) follows because $\sum_{j=1}^m s_j(A) \leq \sqrt{m}\{\sum_{j=1}^m s_j^2(A)\}^{1/2} = \sqrt{m}\|A\|_{\mathrm{F}} = \sqrt{m}$ with $\|A\|_{\mathrm{F}} = 1$ and $\|V\|_{\mathrm{op}} \leq 2L^2(\|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2)$, which can be shown as follows. For any L_2 unit vector w, we have

$$\begin{aligned} &w^{\mathrm{T}}Vw = \mathrm{E}\left[\{w^{\mathrm{T}}(g_1 - g_2)\}^2\right] \leq \mathrm{E}\|g_1 - g_2\|_2^2 \\ &\leq L^2\mathrm{E}\|\mu_1 + D_1Z - (\mu_2 + D_2Z)\|_2^2 \\ &\leq 2L^2\left\{\|\mu_1 - \mu_2\|_2^2 + \mathrm{E}\|(D_1 - D_2)Z\|_2^2\right\} \\ &\leq 2L^2\left(\|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2\right), &(176) \end{aligned}$$

using the fact that $g(\cdot)$ is L-Lipschitz and the marginal variances of Z are 1. The expectation of the second term on (173) can be bounded as

$$\begin{aligned}
&|\mathrm{E}(g_1 - g_2)^{\mathrm{T}} A g_2| \le \mathrm{E}\,|(g_1 - g_2)^{\mathrm{T}} A g_2| \\
&\le \mathrm{E}\|g_1 - g_2\|_2 \|A g_2\|_2 \\
&\le \mathrm{E}^{1/2}(\|g_1 - g_2\|_2^2)\mathrm{E}^{1/2}(\|A g_2\|_2^2) \\
&\le \sqrt{2}L \left(\|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2\right)^{1/2} (\mathrm{E}\|g_2\|_2^2)^{1/2}. \qquad (177)
\end{aligned}$$

Line (177) uses the fact that $\mathrm{E}\{\|g_1 - g_2\|_2^2\} \le 2L^2(\|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2)$ based on (176) and the following argument:

$$\mathrm{E}\|A g_2\|_2^2 \le \mathrm{E}(\|A\|_{\mathrm{op}}^2 \|g_2\|_2^2) \le \mathrm{E}\|g_2\|_2^2,$$

where the last step follows because $\|A\|_{\mathrm{op}} \le \|A\|_{\mathrm{F}} = 1$. Combining (173), (175), and (177) yields the desired result. ■

The following result gives a tighter bound than in Lemma 64 under a boundedness condition in addition to the Lipschitz condition.

Lemma 65 *In the setting of Lemma 64, suppose that each component of $g_1(x)$ and $g_2(x)$ is bounded in $[-1, 1]$. Then for any matrix $A \in \mathbb{R}^{m\times m}$ with $\|A\|_{\mathrm{F}} = 1$,*

$$|\mathrm{E}g^{\mathrm{T}}(X_1) A g(X_1) - \mathrm{E}g^{\mathrm{T}}(X_2) A g(X_2)| \le 2\sqrt{2m} L \Delta,$$

where $\Delta^2 = \|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2$.

Proof The difference $g_1^{\mathrm{T}} A g_1 - g_2^{\mathrm{T}} A g_2$ can also be decomposed as

$$g_1^{\mathrm{T}} A g_1 - g_2^{\mathrm{T}} A g_2 = (g_1 - g_2)^{\mathrm{T}} A g_1 + (g_1 - g_2)^{\mathrm{T}} A g_2.$$

By (177), both of the two terms on the right-hand side can be bounded in absolute values by $\sqrt{2}L\Delta(\mathrm{E}\|g_2\|_2^2)^{1/2}$ and hence by $\sqrt{2m}L\Delta$, because $\mathrm{E}\|g_2\|_2^2 \le m$ by the componentwise boundedness of g_1 and g_2. ■

Appendix E. Technical tools

E.1 von Neumann's trace inequality

Lemma 66 *(Von Neumann, 1937) For any $m \times m$ matrices A and B with singular values $\alpha_1 \ge \cdots \ge \alpha_m \ge 0$ and $\beta_1 \ge \cdots \ge \beta_m \ge 0$ respectively,*

$$|\mathrm{tr}(AB)| \le \sum_{j=1}^{m} \alpha_j \beta_j.$$

As a direct consequence, if A is symmetric and non-negative definite, then

$$|\mathrm{tr}(AB)| \le \mathrm{tr}(A)\|B\|_{\mathrm{op}}$$

This follows because the singular values α_i's are also the eigenvalues of A and hence $\mathrm{tr}(A) = \sum_{j=1}^{m} \alpha_j$ for a symmetric and nonnegative definite matrix A, and $\|B\|_{\mathrm{op}} = \max_{i=j,\ldots,m} \beta_j$ by the definition of $\|B\|_{\mathrm{op}}$.

E.2 Sub-gassisan and sub-exponential properties

The following results can be obtained from Vershynin (2018), Proposition 2.5.2.

Lemma 67 *For a random variable Y, the following properties are equivalent: there exist universal constants $C_{\mathrm{sg},ij} > 0$ such that the $C_{\mathrm{sg},ij}^{-1} K_j \le K_i \le C_{\mathrm{sg},ij} K_j$ for all $1 \le i \ne j \le 4$, where K_i is the parameter appearing in property (i).*

(i) $\mathrm{P}(|Y| > t) \le 2\exp(-\frac{t^2}{2K_1^2})$ *for any* $t > 0$.

(ii) $\mathrm{E}^{1/p}(|Y|^p) \le K_2\sqrt{p}$ *for any* $p \ge 1$.

(iii) $\mathrm{E}\exp(Y^2/K_3^2) \le 2$.

If $\mathrm{E}Y = 0$, *then properties (1)–(3) are also equivalent to the following one.*

(iv) $\mathrm{E}\exp(sY) \le \exp(\frac{K_4^2 s^2}{2})$, *for any* $s \in \mathbb{R}$.

Say that Y is a sub-gaussian random variable with tail parameter K if property (1) holds in Lemma 67 with $K_1 = K$. The following result shows that being sub-gaussian depends only on tail probabilities of a random variable.

Lemma 68 *Suppose that for some $b, K > 0$, a random variable Y satisfies that*

$$\mathrm{P}(|Y| > b + t) \le 2\mathrm{e}^{-\frac{t^2}{2K^2}} \quad \text{for any } t > 0.$$

Then Y is sub-gaussian with tail parameter $K + b$.

Proof We distinguish two cases of $y > 0$. First, if $y > K+b$, then $(y-b)/K > y/(K+b) > 1$ and

$$\mathrm{P}(|Y| > y) \le 2\exp\left\{-\frac{(y-b)^2}{2K^2}\right\} \le 2\exp\left\{-\frac{y^2}{2(K+b)^2}\right\}.$$

Second, if $y \le K + b$, then

$$\mathrm{P}(|Y| > y) \le 1 \le 2\mathrm{e}^{-\frac{1}{2}} \le 2\exp\left\{-\frac{y^2}{2(K+b)^2}\right\}.$$

Hence the desired result holds. ■

The following result follows directly from Chernoff's inequality.

Lemma 69 *Suppose that $(Y_1 \ldots, Y_n)$ are independent such that $EY_i = 0$ and Y_i is sub-gaussian with tail parameter K for $i = 1, \ldots, n$. Then*

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| > C_{\mathrm{sg}5} t\right) \le 2\exp\left(-\frac{nt^2}{2K^2}\right) \quad \text{for any } t > 0.$$

where $C_{\mathrm{sg}5} = C_{\mathrm{sg},14}$ as in Lemma 67.

The following result can be obtained from Vershynin (2018), Exercise 2.5.10.

Lemma 70 *Let $(Y_1, \ldots, Y_n)$ be random variables such that Y_i is sub-gaussian with tail parameter K for $i = 1, \ldots, n$. Then*

$$\mathrm{E} \max_{i=1,\ldots,n} |Y_i| \leq C_{\mathrm{sg6}} K \sqrt{\log(2n)},$$

where $C_{\mathrm{sg6}} > 0$ is a universal constant.

Say that $Y \in \mathbb{R}^d$ is a sub-gaussian random vector with tail parameter K if $w^{\mathrm{T}}Y$ is sub-gaussian with tail parameter K for any $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$. The following result can be obtained from Hsu et al. (2012), Theorem 2.1.

Lemma 71 *Suppose that $Y \in \mathbb{R}^d$ with $\mathrm{E}Y = 0$ is a sub-gaussian random vector with tail parameter K. Then for any $t > 0$, we have that with probability at least $1 - \mathrm{e}^{-t}$,*

$$\|Y\|_2 \leq C_{\mathrm{sg7}} K (\sqrt{d} + \sqrt{t}),$$

where $C_{\mathrm{sg7}} > 0$ is a universal constant.

The following result can be obtained from Vershynin (2010), Theorem 5.39 and Remark 5.40(1). Formally, this is different from Vershynin (2018), Theorem 4.7.1 and Exercise 4.7.3, due to assumption (4.24) used in the latter result.

Lemma 72 *Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed as $Y \in \mathbb{R}^d$, where Y is a sub-gaussian random vector with tail parameter K. Then for any $t > 0$, we have that with probability at least $1 - 2\mathrm{e}^{-t}$,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Y_i Y_i^{\mathrm{T}} - \Sigma \right\|_{\mathrm{op}} \leq C_{\mathrm{sg8}} K^2 \left(\sqrt{\frac{d+t}{n}} + \frac{d+t}{n} \right),$$

where $\Sigma = \mathrm{E}(YY^{\mathrm{T}})$ and C_{sg8} is a universal constant.

The following result can be obtained from Vershynin (2018), Proposition 2.5.2.

Lemma 73 *For a random variable Y, the following properties are equivalent: there exist universal constants $C_{\mathrm{sx},ij} > 0$ such that the $C_{\mathrm{sx},ij}^{-1} K_j \leq K_i \leq C_{\mathrm{sx,ij}} K_j$ for all $1 \leq i \neq j \leq 4$, where K_i is the parameter appearing in property (i).*

(i) $\mathrm{P}(|Y| > t) \leq 2\exp(-\frac{t}{K_1})$ *for any $t > 0$.*

(ii) $\mathrm{E}^{1/p}(|Y|^p) \leq K_2 p$ *for any $p \geq 1$.*

(iii) $\mathrm{E}\exp(|Y|/K_3) \leq 2$.

If $\mathrm{E}Y = 0$, then properties (i)–(iii) are also equivalent to the following one.

(iv) $\mathrm{E}\exp(sY) \leq \exp(\frac{K_4^2 s^2}{2})$, *for any $s \in \mathbb{R}$ satisfying $|s| \leq K_4^{-1}$.*

Say that Y is a sub-exponential random variable with tail parameter K if property (1) holds in Lemma 73 with $K_1 = K$. The following result, from Vershynin (2018), Lemma 2.7.7, provides a link from sub-gaussian to sub-exponential random variables.

Lemma 74 *Suppose that Y_1 and Y_2 are sub-gaussian random variables with tail parameters K_1 and K_2 respectively. Then $Y_1 Y_2$ is sub-exponential with tail parameter $C_{\text{sx5}} K_1 K_2$, where $C_{\text{sx5}} > 0$ is a universal constant.*

The following result about centering can be obtained from Vershynin (2018), Exercise 2.7.10.

Lemma 75 *Suppose that Y is sub-exponential random variable with tail parameter K. Then $Y - \mathrm{E}Y$ is sub-exponential random variable with tail parameter $C_{\text{sx6}} K$, where $C_{\text{sx6}} > 0$ is a universal constant.*

The following result can be obtained from Vershynin (2018), Corollary 2.8.3.

Lemma 76 *Suppose that $(Y_1 \ldots, Y_n)$ are independent such that $EY_i = 0$ and Y_i is sub-exponential with tail parameter K for $i = 1, \ldots, n$. Then*

$$\mathrm{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| > C_{\text{sx7}} K \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right)\right\} \leq 2\mathrm{e}^{-t} \quad \text{for any } t > 0.$$

where $C_{\text{sx7}} > 0$ is a universal constant.

E.3 Symmetrization and contraction

The following result can be obtained from the symmetrization inequality (Section 2.3.1 in van der Vaart and Wellner, 1996) and Theorem 7 in Meir and Zhang (2003).

Lemma 77 *Let $X_1, \ldots, X_n$ be i.i.d. random vectors and $\mathcal{F}$ be a class of real-valued functions such that $\mathrm{E}f(X_1) < \infty$ for all $f \in \mathcal{F}$. Then we have*

$$\mathrm{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathrm{E}f(X_i)\right\} \leq 2\mathrm{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right\},$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables that are independent of $X_1, \ldots, X_n$. The above inequality also holds with the left-hand side replaced by

$$\mathrm{E}\sup_{f\in\mathcal{F}}\left\{\mathrm{E}f(X_i) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right\}.$$

Proof For completeness, we give a direct proof. Let $(X_1', \ldots, X_n')$ be i.i.d. copies of $(X_1, \ldots, X_n)$. Then we have

$$\begin{aligned}
&\mathrm{E} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathrm{E} f(X) \right\} \\
&= \mathrm{E}_{X_i} \left[\sup_{f \in \mathcal{F}} \mathrm{E}_{X_i'} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X_i') \right\} \right] \\
&\leq \mathrm{E}_{X_i, X_i'} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(X_i') \right\} \right] \qquad (178) \\
&= \mathrm{E}_{X_i, X_i', \epsilon_i} \left[\sup_{f \in \mathcal{F}} \epsilon_i \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(X_i') \right\} \right] \qquad (179) \\
&\leq \mathrm{E}_{X_i, \epsilon_i} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right\} + \mathrm{E}_{X_i', \epsilon_i} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} -\epsilon_i f(X_i') \right\} \\
&= 2 \mathrm{E}_{X_i, \epsilon_i} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right\}.
\end{aligned}$$

Line (178) follows from Jensen's inequality. Line (179) follows because for a pair of i.i.d. random variables, their difference is a symmetric random variable about 0 and its distribution remains the same when multiplied by an independent Rademacher random variable. A similar argument is also applicable for upper bounding $\mathrm{E} \sup_{f \in \mathcal{F}} \left\{ \mathrm{E} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\}$. ■

Lemma 78 *Let ϕ be a function with a Lipschitz constant R. Then in the setting of Lemma 77, we have*

$$\mathrm{E} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(f(X_i)) \right\} \leq \mathrm{E} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i R f(X_i) \right\}.$$

E.4 Entropy and maximal inequality

For a function class $\mathcal{F}$ in a metric space endowed with norm $\|\cdot\|$, the covering number $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|)$ is defined as the smallest number of balls of radius δ in the $\|\cdot\|$-metric needed to cover $\mathcal{F}$. The entropy, $H(\delta, \mathcal{F}, \|\cdot\|)$ is defined as $\log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|)$. The following maximal inequality can be obtained from Dudley's inequality for sub-gaussian variables (e.g., van de Geer, 2000, Corollary 8.3; Bellec et al., 2018, Proposition 9.2) including Rademacher variables.

Lemma 79 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$, and $(\epsilon_1, \ldots, \epsilon_n)$ be independent Rademacher random variables. For a fixed set of points $\{x_i \in \mathcal{X} : i = 1, \ldots, n\}$, define the random variable*

$$Z_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right|.$$

Suppose that $\sup_{f\in\mathcal{F}}\|f\|_n \le 1$ *and* $\int_0^1 H^{1/2}(u,\mathcal{F},\|\cdot\|_n)\,\mathrm{d}u \le \Psi_n(\mathcal{F})$*, where* $\|\cdot\|_n$ *is the empirical* L_2 *norm,* $\|f\|_n = \{n^{-1}\sum_{i=1}^n f^2(x_i)\}^{1/2}$*. Then for any* $t>0$*,*

$$\mathrm{P}\left\{Z_n(\mathcal{F})/C_{\mathrm{rad}} > n^{-1/2}(\Psi_n(\mathcal{F})+t)\right\} \le 2\mathrm{e}^{-\frac{t^2}{2}}, \tag{180}$$

where $C_{\mathrm{rad}}>0$ *is a universal constant.*

The following result, taken from van der Vaart and Wellner (1996), Theorem 2.6.7, provides an upper bound on the entropy of a function class in terms of the VC index. For any $r\ge 1$ and probability measure Q, the $L_r(Q)$ norm is defined as $\|f\|_{r,Q} = (\int |f|^r\,\mathrm{d}Q)^{1/r}$.

Lemma 80 *Let* $\mathcal{F}$ *be a VC class of functions such that* $\sup_{f\in\mathcal{F}}|f|\le 1$*. Then for any* $r\ge 1$ *and probability measure* Q*, we have*

$$\mathcal{N}(u,\mathcal{F},\|\cdot\|_{r,Q}) \le C_{\mathrm{vc}}V(\mathcal{F})(16\mathrm{e})^{V(\mathcal{F})}u^{-r(V(\mathcal{F})-1)} \quad \text{for any } u\in(0,1),$$

where $V(\mathcal{F})$ *denotes the VC index of* $\mathcal{F}$ *and* $C_{\mathrm{vc}}\ge 1$ *is a universal constant.*

We deduce the following implications of the preceding results, which can be used in conjunction with Lemmas 60–61.

Corollary 81 *In the setting of Lemma 79, the random variable* $Z_n(\mathcal{F})$ *is sub-gaussian with tail parameter* $C_{\mathrm{rad}}n^{-1/2}(\Psi_n(\mathcal{F})+1)$*.*

Proof By (180), the results follows from an application of Lemma 68. ■

Corollary 82 *In the setting of Lemma 79, the following results hold.*

(i) If $\sup_{f\in\mathcal{F}}|f|\le 1$*, then* $Z_n(\mathcal{F})$ *is sub-gaussian with tail parameter* $C_{\mathrm{rad2}}\sqrt{V(\mathcal{F})/n}$*, where* $C_{\mathrm{rad2}} = C_{\mathrm{rad}}\{1+\sqrt{2+\log(16C_{\mathrm{vc}})}+\int_0^1\sqrt{2\log(u^{-1})}\,\mathrm{d}u\}$*.*

(ii) Consider another two classes $\mathcal{G}$ *and* $\mathcal{H}$ *of functions from* $\mathcal{X}$ *to* $\mathbb{R}$ *in addition to* $\mathcal{F}$*, and let* $\mathcal{F}_{\mathrm{com}} = \{fg+h : f\in\mathcal{F}, g\in\mathcal{G}, h\in\mathcal{H}\}$*. If* $\sup_{f\in\mathcal{F}\cup\mathcal{G}\cup\mathcal{H}}|f|\le 1$*, then* $Z_n(\mathcal{F}_{\mathrm{com}})$ *is sub-gaussian with tail parameter* $C_{\mathrm{rad3}}\sqrt{\{V(\mathcal{F})+V(\mathcal{G})+V(\mathcal{H})\}/n}$*, where* $C_{\mathrm{rad3}} = C_{\mathrm{rad}}\{1+\sqrt{2+\log(16C_{\mathrm{vc}})}+\int_0^1\sqrt{2\log(3u^{-1})}\,\mathrm{d}u\}$*.*

Proof (i) Take $r=2$ and Q to be the empirical distribution on $\{x_1,\ldots,x_n\}$. By Lemma 80, the entropy integral $\int_0^1 H^{1/2}(u,\mathcal{F},\|\cdot\|_n)\,\mathrm{d}u$ can be upper bounded by

$$\begin{aligned}
&\int_0^1 \log^{1/2}\left\{C_{\mathrm{vc}}V(\mathcal{F})(16\mathrm{e})^{V(\mathcal{F})}u^{-2(V(\mathcal{F})-1)}\right\}\,\mathrm{d}u\\
&= \int_0^1 \left\{\log(C_{\mathrm{vc}})+\log V(\mathcal{F})+V(\mathcal{F})\log(16\mathrm{e})+2(V(\mathcal{F})-1)\log(u^{-1})\right\}^{1/2}\,\mathrm{d}u\\
&\le \sqrt{V(\mathcal{F})}\int_0^1\left\{\sqrt{2+\log(16C_{\mathrm{vc}})}+\sqrt{2\log(u^{-1})}\right\}\,\mathrm{d}u,
\end{aligned}$$

using $\log V(\mathcal{F}) \le V(\mathcal{F})$ for $V(\mathcal{F}) \ge 1$ and $\sqrt{u_1+u_2} \le \sqrt{u_1}+\sqrt{u_2}$. Taking $\Psi_n(\mathcal{F})$ in (i) to be the right-hand side of the preceding display yields the desired result by Corollary 81.

(ii) First, we show that the covering number $\mathcal{N}(u, \mathcal{F}_{\rm com}, \|\cdot\|_n)$ is upper bounded by the product $\mathcal{N}(u/3, \mathcal{F}, \|\cdot\|_n)\mathcal{N}(u/3, \mathcal{G}, \|\cdot\|_n)\mathcal{N}(u/3, \mathcal{H}, \|\cdot\|_n)$. Denote as $\hat{\mathcal{F}}$ a $(u/3)$-net of $\mathcal{F}$ with the cardinality $\mathcal{N}(u/3, \mathcal{F}, \|\cdot\|_n)$. Similarly, denote as $\hat{\mathcal{G}}$ and $\hat{\mathcal{H}}$ those of $\mathcal{G}$, $\mathcal{H}$ with the cardinality $\mathcal{N}(u/3, \mathcal{G}, \|\cdot\|_n)$ and $\mathcal{N}(u/3, \mathcal{H}, \|\cdot\|_n)$ respectively. For any $f \in \mathcal{F}$, $g \in \mathcal{G}$ and $h \in \mathcal{H}$, there exist $\hat{f} \in \hat{\mathcal{F}}$, $\hat{g} \in \hat{\mathcal{G}}$ and $\hat{h} \in \hat{\mathcal{H}}$ such that

$$\|\hat{f} - f\|_n \le u/3, \quad \|\hat{g} - g\|_n \le u/3, \quad \|\hat{h} - h\|_n \le u/3.$$

By the triangle inequality and $\sup_{f \in \mathcal{F} \cup \mathcal{G}} |f| \le 1$, we have

$$\begin{aligned} &\|\hat{f}\hat{g} + \hat{h} - fg - h\|_n \\ &\le \|(\hat{f} - f)\hat{g}\|_n + \|f(\hat{g} - g)\|_n + \|\hat{h} - h\|_n \\ &\le \|\hat{f} - f\|_n + \|\hat{g} - g\|_n + \|\hat{h} - h\|_n \le u. \end{aligned}$$

This shows that $\hat{\mathcal{F}}_{\rm com} = \{\hat{f}\hat{g} + \hat{h} : \hat{f} \in \hat{\mathcal{F}}, \hat{g} \in \hat{\mathcal{G}}, \hat{h} \in \hat{\mathcal{H}}\}$ is a u-net of $\mathcal{F}_{\rm com}$ with respect to $\|\cdot\|_n$. Hence the covering number $\mathcal{N}(u, \mathcal{F}_{\rm com}, \|\cdot\|_n)$ is upper bounded by the cardinality of $\hat{\mathcal{F}}_{\rm com}$, that is, $\mathcal{N}(u/3, \mathcal{F}, \|\cdot\|_n)\mathcal{N}(u/3, \mathcal{G}, \|\cdot\|_n)\mathcal{N}(u/3, \mathcal{H}, \|\cdot\|_n)$.

Next, by Lemma 80 applied to $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ and similar calculation as in (i), the entropy integral $\int_0^1 H^{1/2}(u, \mathcal{F}_{\rm com}, \|\cdot\|_n)\,\mathrm{d}u$ can be upper bounded by

$$\begin{aligned} &\int_0^1 \{\log \mathcal{N}(u/3, \mathcal{F}, \|\cdot\|_n) + \log \mathcal{N}(u/3, \mathcal{G}, \|\cdot\|_n) + \log \mathcal{N}(u/3, \mathcal{H}, \|\cdot\|_n)\}^{1/2}\,\mathrm{d}u \\ &\le \sqrt{V(\mathcal{F}) + V(\mathcal{G}) + V(\mathcal{H})} \int_0^1 \left\{\sqrt{2 + \log(16 C_{\rm vc})} + \sqrt{2\log(3u^{-1})}\right\} \mathrm{d}u. \end{aligned}$$

The desired result follows by applying Corollary 81 to the class $\mathcal{F}_{\rm com}$, with $\Psi_n(\mathcal{F}_{\rm com})$ taken to be the right-hand side of the preceding display. ■

References

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B*, 28:131–142, 1966.

S. Balmand and A. Dalalyan. Convex programming approach to robust estimation of a multivariate Gaussian model. *arXiv preprint arXiv:1512.04734*, 2015.

A. Basu and B. G. Lindsay. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46:683–705, 1994.

A. Basu, I. R. Harris, N. L. Hjort, and M. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85:549–559, 1998.

P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets Lasso: Improved oracle bounds and optimality. *Annals of Statistics*, 46:3603–3642, 2018.

R. Beran. Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, 5:445–463, 1977.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, United Kingdom, 2013.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA, 2005.

R. Butler, P. Davies, and M. Jhun. Asymptotics for the minimum covariance determinant estimator. *Annals of Statistics*, 21:1385–1400, 1993.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Annals of Statistics*, 46:1932–1960, 2018.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48:742–864, 2019.

D. L. Donoho and R. C. Liu. The "automatic" robustness of minimum distance functionals. *Annals of Statistics*, 16:552–586, 1988.

F. Farnia and A. Ozdaglar. Do GANs always have Nash equilibria? In *Proceedings of the 37th International Conference on Machine Learning*, pages 3029–3039, 2020.

H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99:2053–2081, 2008.

C. Gao, J. Liu, Y. Yao, and W. Zhu. Robust estimation and generative adversarial networks. In *International Conference on Learning Representations*, 2019.

C. Gao, Y. Yao, and W. Zhu. Generative adversarial nets for robust scatter estimation: a proper scoring rule perspective. *Journal of Machine Learning Research*, 21:160–1, 2020.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

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

K. Hirose, H. Fujisawa, and J. Sese. Robust sparse Gaussian graphical modeling. *Journal of Multivariate Analysis*, 161:172–190, 2017.

D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley, Hoboken, NJ, 2009.

F. Huszár. An alternative update rule for generative adversarial networks. *Blogpost*, 2016. URL `https://www.inference.vc/an-alternative-update-rule-for-generative-adversarial-networks`.

C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proceedings of the 37th International Conference on Machine Learning*, pages 4880–4889, 2020.

M. Jones, N. L. Hjort, I. R. Harris, and A. Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88:865–873, 2001.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *IEEE 57th Annual Symposium on Foundations of Computer Science*, pages 665–674, 2016.

J. H. Lim and J. C. Ye. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017.

B. G. Lindsay. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Annals of Statistics*, 22:1081–1114, 1994.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Annals of Statistics*, 40:2293–2326, 2012.

Z. Liu and P.-L. Loh. Robust W-GAN-based estimation under Wasserstein contamination. *Information and Inference: A Journal of the IMA*, 12:312–362, 2022.

P.-L. Loh and X. L. Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, 12:1429–1467, 2018.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust Statistics: Theory and Methods*. Wiley, West Sussex, England, 2019.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141: 148–188, 1989.

R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

M. Miyamura and Y. Kano. Robust Gaussian graphical modeling. *Journal of Multivariate Analysis*, 97:1525–1550, 2006.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37:876–904, 2009.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56:5847–5861, 2010.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, 2016.

V. Öllerer and C. Croux. *Robust High-Dimensional Precision Matrix Estimation*, pages 325–350. Springer, Cham, Switzerland, 2015.

D. Paindaveine and G. Van Bever. Halfspace depths for scatter, concentration and shape matrices. *Annals of Statistics*, 46:3276–3307, 2018.

D. M. Rocke. Robustness properties of S-estimators of multivariate location and shape in high dimension. *Annals of Statistics*, 24:1327–1345, 1996.

P. J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 8:37, 1985.

P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283, 1993.

L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801, 1971.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100:441–471, 1987.

R. N. Tamura and D. D. Boos. Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81:223–229, 1986.

Z. Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107:137–158, 2020.

Z. Tan and X. Zhang. On loss functions and regret bounds for multi-category classification. *IEEE Transactions on Information Theory*, 68:5295–5313, 2022.

Z. Tan, Y. Song, and Z. Ou. Calibrated adversarial algorithms for generative modelling. *Stat*, 8:e224, 2019.

G. Tarr, S. Müller, and N. C. Weber. Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis*, 93:404–420, 2016.

J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, pages 523–531, Vancouver, Canada, 1975.

D. E. Tyler. A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, pages 234–251, 1987.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, United Kingdom, 2000.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, NY, 1996.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, United Kingdom, 2018.

J. Von Neumann. Some matrix-inequalities and metrization of matric-space. *Tomsk Univ. Rev.*, 1:286–300, 1937.

Z. Wang and Z. Tan. Tractable and near-optimal adversarial algorithms for robust estimation in contaminated Gaussian models. *arXiv preprint arXiv:2112.12919*, 2021.

M. P. Windham. Robustifying model fitting. *Journal of the Royal Statistical Society: Series B*, 57:599–609, 1995.

K. Wu, G. W. Ding, R. Huang, and Y. Yu. On minimax optimality of GANs for robust mean estimation. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 4541–4551, 2020.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics*, 40:2541–2571, 2012.

J. Zhang. Some extensions of Tukey's depth function. *Journal of Multivariate Analysis*, 82: 134–165, 2002.

J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial networks. In *International Conference on Learning Representations*, 2017.

B. Zhu, J. Jiao, and D. Tse. Deconstructing generative adversarial networks. *IEEE Transactions on Information Theory*, 66:7155–7179, 2020.