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

Ziyue Wang  
Department of Statistics  
Rutgers University  
Piscataway, New Jersey 08854, USA  
zw245@stat.rutgers.edu

Zhiqiang Tan  
Department of Statistics  
Rutgers University  
Piscataway, New Jersey 08854, USA  
ztan@stat.rutgers.edu

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

©2023 Ziyue Wang and Zhiqiang Tan.
License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/ Attribution requirements are provided at http://jmlr.org/papers/v24/22-1053.html
matrix and is of particular interest in Gaussian graphical modeling. In the low-dimensional setting, estimation of $\Sigma^*$ and $\Sigma^*-1$ 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 $\Sigma^*$, 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 $\epsilon$. For various choices of the correlation estimator, such as the transformed Kendall’s $\tau$ and Spearman’s $\rho$ estimator, this method is shown in Loh and Tan (2018) to achieve, in the maximum norm $\|\hat{\Sigma} - \Sigma^*\|_{\text{max}}$, 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 $\Sigma^*$ 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 $\epsilon$.

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 $\|\hat{\Sigma} - \Sigma^*\|_{\text{op}}$, 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 $\beta$-divergences) (Basu et al., 1998; Miyamura and Kano, 2006) and that of $\gamma$-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 \( \epsilon \) 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 \( \epsilon \), 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.
### Table 1: Comparison of existing and proposed methods. OC15, TMW15, LT18 refer to methods and theory in Ollerer 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).

<table>
<thead>
<tr>
<th>Method/Model</th>
<th>Error Rate</th>
<th>Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC15, TMW15, LT18</td>
<td>$\epsilon + \frac{\log(p)}{\sqrt{n}}$</td>
<td>Non-iterative computation</td>
</tr>
<tr>
<td>CGR18</td>
<td>$\epsilon + \sqrt{\frac{p}{n}}$</td>
<td>Minimax optimization with</td>
</tr>
<tr>
<td></td>
<td></td>
<td>zero-one discriminators</td>
</tr>
<tr>
<td>DKKLMS19</td>
<td>$\epsilon$, provided $\epsilon \geq \frac{p}{\sqrt{n}}$</td>
<td>Convex optimization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>up to log factors</td>
</tr>
<tr>
<td>GYZ20</td>
<td>$\epsilon + \sqrt{\frac{p}{n}}$</td>
<td>Minimax optimization with</td>
</tr>
<tr>
<td></td>
<td></td>
<td>neural network discriminators</td>
</tr>
</tbody>
</table>

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 $\chi^2$, and squared Hellinger divergences, but excludes the mixed KL divergence, $\chi^2$ 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 $\epsilon$ 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 $\epsilon$ 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 $\epsilon$ 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 $\epsilon$), 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 $\|a\|_1 = \sum_{i=1}^p |a_i|$, $\|a\|_\infty = \max_{1 \leq i \leq p} |a_i|$, and $\|a\|_2 = (\sum_{i=1}^p a_i^2)^{1/2}$ the $L_1$ norm, $L_\infty$ 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 $\|A\|_{\max} = \max_{1 \leq i \leq m, 1 \leq j \leq n} |a_{ij}|$, the Frobenius norm $\|A\|_F = (\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2)^{1/2}$, the vectorized $L_1$ norm $\|A\|_{1,1} = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|$, the operator norm $\|A\|_{op} = \sup_{\|x\|_2 = 1} \|Ax\|_2$, and the $L_\infty$-induced operator norm $\|A\|_\infty = \sup_{\|x\|_\infty \leq 1} \|Ax\|_\infty$. 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 $\text{vec}(A)$. The cumulative distribution function of the standard normal distribution is denoted by $\Phi(x)$, and the Gaussian error function is denoted by $\text{erf}(x)$.

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), Kendall’s $\tau$ with MAD (Loh and Tan, 2018), and Spearman’s $\rho$ with $Q_n$-estimator (Ollerer 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 $\epsilon$ increases to 20\%. The remaining three methods, Kendall’s $\tau$ with MAD, Spearman’s $\rho$ 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 $\tau$ with MAD and Spearman’s $\rho$ 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 $\tau$ and Spearman’s $\rho$ 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.
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 $P_\theta$ is called a generator and $h_\gamma$ a discriminator. In our study, $P_\theta$ is a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, and $h_\gamma$ 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 $P_\theta$ with density functions $p^*$ and $p_\theta$ is

$$D_f(P^*\|P_\theta) = \int f \left( \frac{p^*(x)}{p_\theta(x)} \right) \, dP_\theta.$$  

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

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

where

$$K_f(P^*, P_\theta; h) = E_{P^*} f'(e^{h(x)}) - E_{P_\theta} f'(e^{h(x)})$$

$$= E_{P^*} f'(e^{h(x)}) - E_{P_\theta} \left\{ e^{h(x)} f'(e^{h(x)}) - f(e^{h(x)}) \right\}.$$
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 $\theta$, it holds that for any function $h$,

$$K_f(P_\gamma, P_\theta; h) \leq D_f(P_\gamma \parallel P_\theta),$$

(2)

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

$$K_{\text{JS}}(P_\gamma, P_\theta; h) = -E_{P_\gamma}\log(1 + e^{-h(x)}) - E_{P_\theta}\log(1 + e^{h(x)}) + \log 4,$$

which is, up to a constant, the expected log-likelihood for logistic regression with log odds function $h(x)$. For $K_f = K_{\text{JS}}$, 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 $D_{r\text{KL}}(P_\gamma \parallel P_\theta) = D_{\text{KL}}(P_\theta \parallel P_\gamma)$ and the objective function

$$K_{r\text{KL}}(P_\gamma, P_\theta; h) = 1 - E_{P_\gamma}e^{-h(x)} - E_{P_\theta}h(x),$$

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

The objective $K_f$ with fixed $\theta$ 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

$$\min_{\theta \in \Theta} \max_{\gamma \in \Gamma} K_{\gamma}(P_\gamma, P_\theta; h_{\gamma}),$$

(3)

where

$$K_{\gamma}(P_\gamma, P_\theta; h) = E_{P_\gamma}\min(1, h(x)) + E_{P_\theta}\min(1, -h(x)).$$

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 $K_{\text{HG}}$ is a nonparametrically tight, lower bound of the total variation distance scaled by 2: for each $\theta$, it holds that for any function $h(x)$,

$$K_{\text{HG}}(P_\gamma, P_\theta; h) \leq 2D_{\text{TV}}(P_\gamma \parallel P_\theta),$$

(4)

where the equality is attained at $h_{\theta\gamma}(x) = \text{sign}(p_\gamma(x) - p_\theta(x))$, and $D_{\text{TV}}(P_\gamma \parallel P_\theta) = \int |p_\gamma(x) - p_\theta(x)|/2\,dx$. The objectives $K_f$ and $K_{\text{HG}}$, with fixed $\theta$, 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
\[
\begin{align*}
\max_{\gamma \in \Gamma} & \quad K_{JS}(P_s, P_\theta; \gamma) \quad \text{with } \theta \text{ fixed}, \\
\min_{\theta \in \Theta} & \quad -E_{P_\theta} \log(1 + e^{-h_\gamma(x)}) \quad \text{with } \gamma \text{ fixed}.
\end{align*}
\]
(5)
The second objective is introduced mainly to overcome vanishing gradients in $\theta$ when the
discriminator is confident. The calibrated rKL-GAN (Huszár, 2016; Tan et al., 2019) is
defined by solving
\[
\begin{align*}
\max_{\gamma \in \Gamma} & \quad K_{JS}(P_s, P_\theta; \gamma) \quad \text{with } \theta \text{ fixed}, \\
\min_{\theta \in \Theta} & \quad -E_{P_\theta} h_\gamma(x) \quad \text{with } \gamma \text{ fixed}.
\end{align*}
\]
(6)
The two objectives are chosen to stabilize gradients in both $\theta$ and $\gamma$ 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
\[
\begin{align*}
\max_{\gamma \in \Gamma} & \quad K_{HG}(P_s, P_\theta; \gamma) \quad \text{with } \theta \text{ fixed}, \\
\min_{\theta \in \Theta} & \quad -E_{P_\theta} h_\gamma(x) \quad \text{with } \gamma \text{ fixed}.
\end{align*}
\]
(7)
Interestingly, the second line in (6) or (7) involves the same objective $-E_{P_\theta} h_\gamma(x)$, which
can be equivalently replaced by $K_{rKL}(P_s, P_\theta; h_\gamma)$ because $\gamma$ and hence $h_\gamma$ 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
\[
\min_{\theta} \max_{\gamma} K(\theta, \gamma),
\]
(8)
if $K(\hat{\theta}, \hat{\gamma}) = \max_{\gamma} K(\hat{\theta}, \gamma) \leq \max_{\theta} K(\theta, \gamma)$ for any $\theta$. In other words, we treat (8) as
nested optimization: $\hat{\theta}$ is a minimizer of $K(\theta, \hat{\gamma}_\theta)$ as a function of $\theta$ and $\hat{\gamma} = \hat{\gamma}_\theta$, where $\hat{\gamma}_\theta$
is a maximizer of $K(\theta, \gamma)$ for fixed $\theta$. 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
\[
\begin{align*}
\max_{\gamma} & \quad K_1(\theta, \gamma) \quad \text{with } \theta \text{ fixed}, \\
\min_{\theta} & \quad K_2(\theta, \gamma) \quad \text{with } \gamma \text{ fixed}.
\end{align*}
\]
(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 $K(\hat{\theta}, \hat{\gamma}) = \max_{\gamma} K(\theta, \gamma) = \min_{\theta} K(\theta, \hat{\gamma})$. 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 $P_x$ replaced by the empirical distribution $P_n$ on real data, initial values $(\theta_0, \gamma_0)$, learning rates $(\alpha_d, \alpha_g)$ for the generator and discriminator, and number of epochs $T$.

for $t = 1 \ldots T$ 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 $\theta_t = \theta_{t-1} - \alpha_g \nabla_\theta K(\theta, \gamma)|_{\theta_{t-1}}$. |

end

Algorithm 2: Gradient descent with concave inner optimization

**Require** A GAN objective function $K(\theta, \gamma)$, which is concave in $\gamma$ for each fixed $\theta$, initial value $\theta_0$, learning rate $\alpha_g$ for the generator, and number of epochs $T$.

for $t = 1 \ldots T$ do

| Sampling: | Generate a sample of fake data to approximate $P_{\theta_{t-1}}$. |
| Updating: | Compute $\gamma_t = \arg\max_\gamma K(\theta_{t-1}, \gamma)$ by a concave optimizer, and $\theta_t = \theta_{t-1} - \alpha_g \nabla_\theta K(\theta, \gamma)|_{\theta_{t-1}}$. |

end

they coincide and yield a Nash equilibrium by Sion’s minimax theorem in the special setting where $K(\theta, \gamma)$ is convex in $\theta$ for each $\gamma$ and concave in $\gamma$ for each $\theta$. 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

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

for any $\theta$ satisfying $\|\theta - \hat{\theta}\| \leq \delta$. 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 $\theta$ and $\gamma$). 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 $\gamma$ for each fixed $\theta$. 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 $\gamma$ to find a local minimizer of $\min_{K}$ for our methods with concave $K$. The population $\min_{K}$ is always a local minimax point, and hence finding a local minimax point is a stationary point. Nevertheless, our GAN methods in Section 4 are designed such that $\min_{K}$ is always a local minimax point, and for which a performance guarantee is derived in Jin et al. (2020). Based on these observations, our GAN methods with concave $K$ are numerically more tractable than nonconvex-nonconcave GANs.

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

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

where $f^{*}$ is the Fenchel conjugate of $f$, i.e., $f^{*}(s) = \sup_{t \in (0, \infty)} (st - f(t))$ and $T_{\gamma}$ is a function taking values in the domain of $f^{*}$. Typically, $T_{\gamma}$ 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 $\tau_{f}(u) = f'(e^{u})$ 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 $P_{\ast}$ and $P_{0}$.

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

$$\min_{\theta \in \Theta} \max_{\gamma \in \Gamma} \{ E_{\theta} \text{sign}(h_{\gamma}(x)) - E_{\theta} \text{sign}(h_{\gamma}(x)) \}. \quad (11)$$

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 $f(t) = \max(1, t)$, the conjugate is $f^{*}(s) = \max(-1, s)$ if $s \leq 1$ or $\infty$ if $s > 1$. If $T_{\gamma}$ is specified as $T_{\gamma} = \min(1, h_{\gamma}(x))$, then the objective in $f$-GAN (10) can be shown to be

$$E_{P_{\ast}} \min(1, h_{\gamma}(x)) + E_{P_{0}} \min(1, \max(-1, -h_{\gamma}(x))),$$

which in general differs from the negative hinge loss in (3) unless $h_{\gamma}$ is upper bounded by 1. If $h_{\gamma}$ is specified as $2 \text{sigmoid}(h_{\gamma}) - 1$ 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

$$\min_{\theta \in \Theta} \max_{\gamma \in \Gamma} \{ E_{\theta} \text{sigmoid}(\tilde{h}_{\gamma}(x)) - E_{\theta} \text{sigmoid}(\tilde{h}_{\gamma}(x)) \}. \quad (12)$$

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 $\epsilon$-contamination model, that is, the data distribution $P_\epsilon$ is of the form

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

where $P_{\theta^*}$ is $N(\mu^*, \Sigma^*)$ with unknown $\theta^* = (\mu^*, \Sigma^*)$, $Q$ is a probability distribution for contaminated data, and $\epsilon$ is a contamination fraction. Both $Q$ and $\epsilon$ are unknown and $Q$ can be an arbitrary probability distribution. The dependency of $P_\epsilon$ on $(\theta^*, Q)$ 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 $P(U_i = 1) = \epsilon$ and $X_i$ is drawn from $P_{\theta^*}$ 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 $\Theta_1 = \{ (\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma \succeq 0, \|\Sigma\|_{\text{max}} \leq M_1 \}$ for a constant $M_1 > 0$. Equivalently, the diagonal elements of $\Sigma$ are upper bounded by $M_1$ for $(\mu, \Sigma) \in \Theta_1$. The second choice is $\Theta_2 = \{ (\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma \succeq 0, \|\Sigma\|_{\text{op}} \leq M_2 \}$ for a constant $M_2 > 0$. For simplicity, the dependency of $\Theta_1$ on $M_1$ or $\Theta_2$ on $M_2$ is suppressed in the notation. For the second parameter space $\Theta_2$, 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 $\Theta_1$, 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 $\Theta_2$.

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

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 $\max_{\gamma \in \Gamma} K_f(P_\epsilon, P_{\theta^*}; h_\gamma) = D_f(P_\epsilon || P_{\theta^*})$ for each $\theta$, then the (population) logit $f$-GAN amounts to minimizing the $f$-divergence $D_f(P_\epsilon || P_{\theta^*})$. Similarly, if the discriminator class $\{h_\gamma \}$ in (3) is sufficiently rich, then the (population) hinge GAN amounts to minimizing the total variation $D_{\text{TV}}(P_\epsilon || P_{\theta^*})$.

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

\( f \)-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 \( f : (0, \infty) \rightarrow \mathbb{R} \) 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 \( \Theta_0 = \{ (\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma \text{ is a } p \times p \text{ variance matrix} \} \).

(i) Assume that \( f \) satisfies Assumption 1. Let \( \hat{\theta} = \arg\min_{\theta \in \Theta_0} D_f(p_\theta || p_\theta) \). If \( \sqrt{-2(f''(1))^{-1}f'(1/2)} \epsilon + \epsilon < 1/2 \), then for any contamination distribution \( Q \),

\[
\|\hat{\mu} - \mu^*\|_2 \leq C\|\Sigma^*\|_{\text{op}}^{1/2}\sqrt{\epsilon}, \quad \|\hat{\mu} - \mu^*\|_{\infty} \leq C\|\Sigma^*\|_{\text{max}}^{1/2}\sqrt{\epsilon},
\]

(14)

and

\[
\|\hat{\Sigma} - \Sigma^*\|_{\text{op}} \leq C\|\Sigma^*\|_{\text{op}}\sqrt{\epsilon}, \quad \|\hat{\Sigma} - \Sigma^*\|_{\text{max}} \leq C\|\Sigma^*\|_{\text{max}}\sqrt{\epsilon},
\]

(15)

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

(ii) Let \( \hat{\theta} = \arg\min_{\theta \in \Theta_0} D_{\text{TV}}(p_\theta || p_\theta) \). If \( \epsilon < 1/4 \) then (14) and (15) hold for an absolute constant \( C > 0 \) with \( \sqrt{\epsilon} \) replaced by \( \epsilon \) throughout.
Figure 2: Illustration of robustness of minimum $f$-divergence location estimation. Figure (a): Location error $|\bar{\mu} - \mu^*|$ against contamination fraction $\epsilon$ from 0 to 0.015, with $P_\theta$ being $N(0, 1)$ and contamination $Q$ being Cauchy$(5, 1/2)$ fixed; Figure (b): Location error $|\bar{\mu} - \mu^*|$ against contamination location $\mu_Q$ from 0 to 10, with $P_\theta$ being $N(0, 1)$, contamination fraction $\epsilon = 0.1$ fixed, and contamination $Q$ being Cauchy$(\mu_Q, 1/2)$. The squared Hellinger, reverse $\chi^2$, 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 $\chi^2$) 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 $\epsilon$, for $\epsilon$ close to 0. For the KL, mixed KL, and $\chi^2$ 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 $\text{Err}_{f_0}(\epsilon)$, 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 \( \hat{\theta} = (\hat{\mu}, \hat{\Sigma}) \) is defined by minimization over the unrestricted space \( \Theta_0 \) instead of \( \Theta_1 \) or \( \Theta_2 \) with the restriction \( \|\Sigma\|_{\text{max}} \leq M_1 \) or \( \|\Sigma\|_{\text{op}} \leq M_2 \). The population bounds are also adaptive in that the scaling constants depend directly on the maximum or operator norm of the true variance matrix \( \Sigma^* \), instead of pre-specified constants \( M_1 \) or \( M_2 \). Note that the parameter space is also restricted such that \( \|\Sigma\|_{\text{op}} \leq M_2 \) 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 \( \|\Sigma^*\|_{\text{max}} \), which can be much smaller than \( \|\Sigma^*\|_{\text{op}} \) 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 \( D(P, P_0) \) satisfying the triangle inequality. For minimum TV estimation, let \( \theta_P = (\mu_P, \Sigma_P) = \arg\min_{\theta} D_{TV}(P||P_0) \). For location estimation, define

\[
\begin{align*}
    b(\epsilon) &= \sup_{P, D_{TV}(P||P_0) \leq \epsilon} \|\mu_P - \mu^*\|_2, \\
    b_0(\epsilon) &= \sup_{\theta, D_{TV}(P_0||P_{0\epsilon}) \leq \epsilon} \|\mu - \mu^*\|_2,
\end{align*}
\]

which are called the bias distortion curve and the gauge function. Scatter estimation can be discussed in a similar manner. For a general family \( \{P_0\} \), the first half in our proof of Theorem 6(ii) shows that for any \( P \) satisfying \( D_{TV}(P||P_{0\epsilon}) \leq \epsilon \), we have \( D_{TV}(P_{\theta_P}, P_{0\epsilon}) \leq 2\epsilon \). This implies a bound similar to Proposition 5.1 in Donoho and Liu (1988):

\[
b(\epsilon) \leq b_0(2\epsilon). \tag{16}
\]

For the multivariate Gaussian family \( \{P_0\} \), the second half in our proof of Theorem 6(ii) derives an explicit upper bound on \( b_0(\epsilon) \) provided that \( 2\epsilon \leq a \) for a constant \( a \in [0, 1/2] \):

\[
b_0(2\epsilon) \leq S_{1,a} \|\Sigma^*\|_{\text{op}}^{1/2}(2\epsilon),
\]

where \( S_{1,a} = \{\Phi(\Phi^{-1}(1/2 + a))\}^{-1} \). Combining the preceding inequalities yields \( b(\epsilon) \leq C\|\Sigma^*\|_{\text{op}}^{1/2}\epsilon \) 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 \( D(P, P_0) \), including the Hellinger distance, where the MD functional \( \theta_P = (\mu_P, \Sigma_P) \) is defined as \( \arg\min_{\theta} D=P_{\theta P} \), and \( b(\epsilon) \) and \( b_0(\epsilon) \) are defined with \( D_{TV}(P||P_{0\epsilon}) \) replaced by \( D(P, P_0) \). 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 \( \mu^* \), the population breakdown point of a functional \( T = T(P) \) can be defined as \( \sup \{\epsilon : b_T(\epsilon) < \infty\} \), where \( b_T(\epsilon) = \sup_{P, D_{TV}(P||P_{0\epsilon}) \leq \epsilon} \|T(P) - \mu^*\|_2 \). 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 \( b_T(\epsilon) \leq C\|\Sigma^*\|_{\text{op}}^{1/2}\epsilon \), 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 \( b_T(\epsilon) \leq C\|\Sigma^*\|_{op}^{1/2} \sqrt{\epsilon} \), 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 \( \epsilon \) and some scaling constants free of \( \epsilon \).

### 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 \((\epsilon, Q)\). 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 \( \mathcal{H}_{sp} \):

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

where \( \gamma = (\gamma_0, \gamma_1^T, \gamma_2^T)^T \in \Gamma \) with \( \Gamma = \mathbb{R}^{l+5p+(5p)^2} \) and \( \varphi(x) = (\varphi_1^T(x), \ldots, \varphi_5^T(x))^T \). The basis vector \( \varphi_t(x) \in \mathbb{R}^p \) is obtained by applying \( t \mapsto (t - \xi_t)_+ \) componentwise to \( x = (x_1, \ldots, x_p)^T \), with the fixed knot \( \xi_t = -2, -1, 0, 1, \) or 2 for \( t = 1, \ldots, 5 \) respectively. For concreteness, assume that every two components of \( \gamma_2 \) are identical if associated with the same product of two components of \( \varphi(x) \), that is, \( \gamma_2 \) 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 \( \mathcal{H}_{sp} \) 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 \( 1 \leq j_1, j_2 \leq p \). 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 \( \Theta_1 \) or \( \Theta_2 \) 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 \( \chi^2 \) 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 $\gamma$. For $j = 1, \ldots, p$, $\phi^{(j)}$ is a vector of truncated linear (TL) basis functions of $x_j$ (the $j$th 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) $f'(e^u)$ is concave in $u \in \mathbb{R}$.

(ii) $f''(e^u)$ 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

$$\min_{\theta \in \Theta} \max_{\gamma \in \Gamma} \{ K_f(P_n, P_\theta; h_{\gamma,\mu}) - \lambda_1 \text{pen}_1(\gamma) \},$$

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 $\{X_1, \ldots, X_n\}$, $h_{\gamma,\mu}(x) = h_{\text{sp},\gamma}(x - \mu)$, $\text{pen}_1(\gamma) = \|\gamma_1\|_1 + \|\gamma_2\|_1$, the $L_1$ norm of $\gamma$ excluding the intercept $\gamma_0$, 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 $\gamma$, to achieve suitable control of sampling variation. Second, the discriminator $h_{\gamma,\mu}$ is a spline function with knots depending on $\mu$, the location parameter for the generator. By a change of variables, the non-penalized objective in (17) can be equivalently written as

$$K_f(P_n, P_\theta; h_{\gamma,\mu}) = E_{P_n - \mu} f'(e^{h_{\text{sp},\gamma}(x)}) - E_{P_0, \Sigma} f''(e^{h_{\text{sp},\gamma}(x)}),$$

where $K_f(P_n, P_\theta; h)$ is the $f$-divergence between $P_n$ and $P_\theta$. The diagram shows the connection between the input variables $x_1, x_2, \ldots, x_p$ and the output $h_{\text{sp},\gamma}(x)$ through the truncated linear basis functions and pairwise interactions.
where $P_n - \mu$ denotes the empirical distribution on $\{X_1 - \mu, \ldots, X_n - \mu\}$. Hence $K_f(P_n, P_0; h_{\gamma, \mu})$ 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 $(n, p, \epsilon, \delta)$ and $Q$ under the stated conditions.

**Theorem 11** 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). For $\delta < 1/7$, if $\lambda_1 \geq C_1 \left(\sqrt{\log p/n} + \sqrt{\log(1/\delta)/n}\right)$ and $\sqrt{\rho + 1/(n\delta)} + \lambda_1 \leq C_2$, then with probability at least $1 - 7\delta$ the following bounds hold uniformly over contamination distribution $Q$,

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

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

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

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 $\sqrt{\rho} + \sqrt{\log(p)/n}$. 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 $\sqrt{\rho}$ 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

$$\min_{\theta \in \Theta_2} \max_{\gamma \in \Gamma} \{K_f(P_n, P_0; h_{\gamma, \mu}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2)\}, \quad (19)$$

where $K_f(P_n, P_0; h)$ and $h_{\gamma, \mu}(x)$ are defined as in (17), $\text{pen}_2(\gamma_1) = \|\gamma_1\|_2$ and $\text{pen}_2(\gamma_2) = \|\gamma_2\|_2$, the $L_2$ norms of $\gamma_1$ and $\gamma_2$, and $\lambda_2 \geq 0$ and $\lambda_3 \geq 0$ are tuning parameters. Compared with $L_1$ penalized logit $f$-GAN (17), the $L_2$ norms of $\gamma_1$ and $\gamma_2$ are separately associated with tuning parameters $\lambda_2$ and $\lambda_3$ 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 $\lambda_2$ and $\lambda_3$ is crucial for achieving meaningful error bounds in the $L_2$ and Frobenius norms for simultaneous estimation of $(\mu^*, \Sigma^*)$. 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 $\|\Sigma^*\|_{\op} \leq M_2$, $f$ satisfies Assumptions 1–2, and $pe$ is upper bounded by a constant $B$. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (19). For $\delta < 1/8$, if $\lambda_2 \geq C_1 \left(\sqrt{p/n} + \sqrt{\log(1/\delta)/n}\right)$, $\lambda_3 \geq C_1 \sqrt{B} \left(\sqrt{p/n} + \sqrt{\log(1/\delta)/n}\right)$, and $\sqrt{\rho} + \sqrt{1/(n\delta)} + \lambda_2 \leq C_2$, then with probability at least $1 - 8\delta$ the following bounds hold uniformly over
contamination distribution $Q$,
\[
\|\hat{\mu} - \mu^*\|_2 \leq C \left( \sqrt{\epsilon} + \sqrt{1/(n\delta)} + \lambda_2 \right),
\]
\[
p^{-1/2}\|\hat{\Sigma} - \Sigma^*\|_F \leq C \left( \sqrt{\epsilon} + \sqrt{1/(n\delta)} + \lambda_2 + \lambda_3/\sqrt{p} \right),
\]
where $C_1, C_2, C > 0$ are constants, depending on $M_2$ and $f$ but independent of $(n, p, \epsilon, \delta)$ except through the bound $B$ on $pe$.

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 $pe$ 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 and non-increasing $f$, concavity of $f' (e^u)$ 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 $g(u) = f'(e^u)$, 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 $f'(e^u)$ in $u$, as shown by the Pearson $\chi^2$ in Table 2. Second, for convex and non-increasing $f$, concavity of $f'(e^u)$ in $u \in \mathbb{R}$ also implies concavity of $-f^\#(e^u)$ 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, $-f^\#(e^u) = g(f'(e^u))$ is concave, where $g = -f^*$ is concave and non-decreasing by non-increasingness of $f$.

**Remark 14** The concavity of $f'(e^u)$ and $-f^\#(e^u)$ in $u$ from Assumptions 1(iii) 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 $f'(e^u)$ and $-f^\#(e^u)$ in $u$, in conjunction with the linearity of the spline discriminator $h_{\gamma, \mu}$ in $\gamma$, indicates that the objective function $K_f(P_n, P_\theta; h_{\gamma, \mu})$ is concave in $\gamma$ for any fixed $\theta$. 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 $(\epsilon, p)$ over the parameter space $\Theta_1$ or $\Theta_2$ 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
\[
\min_{\theta \in \Theta_1} \max_{\gamma \in \Gamma} \{ K_{HG}(P_n, P_0; h_{\gamma, \mu}) - \lambda_1 \text{pen}_1(\gamma) \},
\] (20)
where $K_{HG}(P_n, P_0; h)$ is the hinge objective $K_{HG}(P_*, P_0; h)$ in (3) with $P_*$ replaced by $P_n$ and, similarly as in $L_1$ penalized logit $f$-GAN (17), $h_{\gamma, \mu}(x) = h_{sp, \gamma}(x - \mu)$, $\text{pen}_1(\gamma) = \|\gamma_1\|_1 + \|\gamma_2\|_1$, 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 $\lambda_2 \geq C_1 \Big( \sqrt{\log p/n + \log (1/\delta)/n} \Big)$ 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 $(n, p, \epsilon, \delta)$.

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 $\sqrt{\epsilon + \log p/n}$ in terms of dependency on $\epsilon$ 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 $\Theta_1$.

The second method, $L_2$ penalized hinge GAN, is defined by solving
\[
\min_{\theta \in \Theta_2} \max_{\gamma \in \Gamma} \{ K_{HG}(P_n, P_0; h_{\gamma, \mu}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \},
\] (21)
where, similarly as in $L_2$ penalized logit $f$-GAN (19), $h_{\gamma, \mu}(x) = h_{sp, \gamma}(x - \mu)$, $\text{pen}_2(\gamma_1) = \|\gamma_1\|_2$ and $\text{pen}_2(\gamma_2) = \|\gamma_2\|_2$, and $\lambda_2 \geq 0$ and $\lambda_3 \geq 0$ are tuning parameters.

**Theorem 16** Assume that $\|\Sigma^*\|_{op} \leq M_2$. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (21). For $\delta < 1/8$, if $\lambda_2 \geq C_1 \Big( \sqrt{p/n + \log (1/\delta)/n} \Big)$, $\lambda_3 \geq C_1 \sqrt{p} \Big( \sqrt{p/n + \log (1/\delta)/n} \Big)$, 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 $Q$,
\[
\|\hat{\mu} - \mu^*\|_2 \leq C \left( \sqrt{p} \left( \epsilon + \sqrt{\epsilon/(n\delta)} \right) + \lambda_2 \right),
\]
\[
p^{-1/2}\|\hat{\Sigma} - \Sigma^*\|_F \leq C \left( \sqrt{p} \left( \epsilon + \sqrt{\epsilon/(n\delta)} \right) + \lambda_2 + \lambda_3/\sqrt{p} \right),
\]
where $C_1, C_2, C > 0$ are constants, depending on $M_2$ but independent of $(n, p, \epsilon, \delta)$. 


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 $\Theta_2$ 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 $K_{HG}(P_n, P_\theta; h)$. This is reminiscent of the concavity of $f'(e^h)$ and $-f'(e^h)$ 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 $K_{HG}(P_n, P_\theta; h)$ in $h$, together with the linearity of the spline discriminator $h_{\gamma, \mu}$ in $\gamma$, implies that the objective function $K_{HG}(P_n, P_\theta; h_{\gamma, \mu})$ is concave in $\gamma$ for any fixed $\theta$. 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):

$$
\begin{align*}
\max_{\gamma \in \Gamma} K_f(P_n, P_\theta; h_{\gamma, \mu}) &- \text{pen}(\gamma; \lambda) \\
\min_{\theta \in \Theta} \mathbb{E}_{P_n} f'(e^{h_{\gamma, \mu}(x)}) - \mathbb{E}_{P_\theta} G(h_{\gamma, \mu}(x)) & \quad \text{with } \gamma \text{ fixed.}
\end{align*}
$$

(22)

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

$$
\begin{align*}
\max_{\gamma \in \Gamma} K_{HG}(P_n, P_\theta; h_{\gamma, \mu}) &- \text{pen}(\gamma; \lambda) \\
\min_{\theta \in \Theta} \mathbb{E}_{P_n} \min(h_{\gamma, \mu}(x), 1) - \mathbb{E}_{P_\theta} G(h_{\gamma, \mu}(x)) & \quad \text{with } \gamma \text{ fixed.}
\end{align*}
$$

(23)

Here $\text{pen}(\gamma; \lambda)$ is an $L_1$ penalty, $\lambda_1(\|\gamma_1\|_1 + \|\gamma_2\|_1)$ and $\Theta_1 = \{ (\mu, \Sigma) : \mu \in \mathbb{R}^p, \|\Sigma\|_{\max} \leq M_1 \}$ as in (17), or $\text{pen}(\gamma; \lambda)$ is an $L_2$ penalty $\lambda_2\|\gamma_1\|_2 + \lambda_3\|\gamma_2\|_2$ and $\Theta_2 = \{ (\mu, \Sigma) : \mu \in \mathbb{R}^p, \|\Sigma\|_{\op} \leq M_2 \}$ 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 $\mu$, so that $\mathbb{E}_{P_n} f'(e^{h_{\gamma, \mu}(x)})$
cannot be dropped in the optimization over $\theta$ 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 $\gamma$ as in (17) or (19) and the corresponding choice of $\Theta$, 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 $\gamma$ as in (20) or (21) and the corresponding choice of $\Theta$, 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 $\mu$ to facilitate theoretical analysis as mentioned in Section 4.2. If $h_{\gamma, \mu}$ were replaced by a discriminator $h_{\gamma}$ defined independently of $\theta$, then taking $K_f = K_{\text{IS}}$ and $G(h) = -\log(1 + e^{-h})$ or $G(h) = -h$ in (22) reduces to GAN with logD trick (5) or calibrated rKL-GAN (6) respectively, and taking $K_f = K_{\text{HG}}$ 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

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

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

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

where $K_f$ is the logit $f$-GAN objective as in (17) and (19), $h_{\gamma, \mu, \Sigma}(x) = h_{\text{sp, } \gamma}(\Sigma^{-1/2}(x - \mu))$ and $\text{pen}(\gamma; \lambda)$ is an $L_1$ or $L_2$ penalty term. The discriminator $h_{\gamma, \mu, \Sigma}(x)$ is obtained by applying $h_{\text{sp, } \gamma}(\cdot)$ 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

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

22
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 $(\mu, \Sigma)$, while employing the spline discriminators $h_{\text{sp}, \gamma}(x)$ 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 $(\mu, \Sigma)$ 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 version of the GAN studied in Gao et al. (2020) is defined as

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

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

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

The objective $L_g(P_*, P_\theta; q)$ is shown to be a lower bound, being tight if $q = 2dP_* / d(P_* + P_\theta)$, for the divergence $D_{g_0}(P_\theta \parallel (P_* + P_\theta) / 2)$, 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 $q_\gamma(x) = \text{sigmoid}(h_\gamma(x))$. Then $L_g(P_*, P_\theta; q_\gamma) = K_f(P_*, P_\theta; h_\gamma)$ for $K_f$ defined in (1) and $f(t) = \frac{1 + t}{2}g_0\left(\frac{2t}{1 + t}\right)$ satisfying that $D_{g_0}(P_\theta \parallel (P_* + P_\theta) / 2) = D_f(P_* \parallel P_\theta)$. 

23
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

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

(28)

where $(\gamma_j^{(2)}, \gamma_j^{(2)})$, $j = 1, \ldots, J$, 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 $\sum_{j=1}^J |\gamma_j^{(1)}| \leq \kappa$ for a tuning parameter $\kappa$. Assume that $g(\eta)$ is three-times continuously differentiable at $\eta = 1/2$, $g''(1/2) > 0$, and for a universal constant $c_0 > 0$,

$$
2g''(1/2) \geq g'''(1/2) + c_0,
$$

(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 $L_g(P_n, P_0; q_\gamma)$ may exhibit a complex, non-concave landscape in $\gamma$, 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 $\frac{d^2}{u^2} f'(e^u)|_{u=0} \geq c_0$ 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 $\sum_{j=1}^J |\gamma_j^{(1)}| \leq \kappa$, the log-odds discriminator $h_\gamma(x)$ in (28) is forced to be bounded, $|h_\gamma(x)| \leq \kappa$, where $\kappa$ 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 $L_g(P_\ast, P_0; q_\gamma)$ over such constrained discriminators may produce a divergence with a substantial gap to the actual divergence $D_f(P_\ast || P_0)$ for any fixed $\theta$. Instead, the implied divergence measure may behave more similarly as the total variation $D_{\text{TV}}(P_\ast || P_0)$ than as $D_f(P_\ast || P_0)$, due to the boundedness of $h_\gamma(x)$ by a sufficiently small $\kappa$, 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 $Q$, $\epsilon$, $n$, and $p$. Results about error dependency on $\epsilon$ 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 \((\mu_0, \Sigma_0)\) 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 \( \gamma \) 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 \( \gamma \). 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 \( \lambda \) 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 $p-2p-[p/2]-1$ 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` (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 $\Sigma^\ast$, 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 $N(0, \Sigma^\ast)$ where $\Sigma^\ast$ 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 $\tau$, and Spearman’s $\rho$. 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 $1_p$.

- $Q = \text{Cauchy} \left( 2.25\epsilon, \frac{1}{3}I_p \right)$ where $c = (1, -1, 1, -1, 1, \ldots)$ is a $p$-dimensional vector of alternating $\pm 1$. 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 $\epsilon$ 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 $\epsilon$ increases to 20%, especially with the first contamination. The pairwise methods, Kendall’s $\tau$ with MAD and Spearman’s $\rho$ 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 $\epsilon$ increases. However, the dependency on $\epsilon$ 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 $\epsilon$ than in the worst scenario. For further understanding, we present in Figure 8 (Appendix A.5) a comparison between two types of contamination...
Table 3: Comparison of existing methods and proposed $L_1$ penalized GAN methods ($p = 100$, $n = 20000$, and varying $\epsilon$ from 0% to 20%). Estimation error of the variance matrix is reported in the maximum norm $\| \cdot \|_{\text{max}}$.

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

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

(i) Assume that \( f \) satisfies Assumption 1, and \( \epsilon \in [0, \epsilon_0] \) for a constant \( \epsilon_0 \in [0, 1/2) \). Let \( \bar{\theta} = \arg\min_{\theta \in \Theta_0} D_f(P_\theta \| P_\theta) \). If \( \text{Err}_{10}(\epsilon) \leq a \) for a constant \( a \in [0, 1/2) \), then we have

\[
\| \bar{\mu} - \mu^* \|_2 \leq S_{1,a} \| \Sigma^* \|_{op}^{1/2} \text{Err}_{f0}(\epsilon),
\]

\[
\| \bar{\mu} - \mu^* \|_{\infty} \leq S_{1,a} \| \Sigma^* \|_{\max}^{1/2} \text{Err}_{f0}(\epsilon),
\]

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

\[
\| \bar{\Sigma} - \Sigma^* \|_{op} \leq 2S_{3,a} \| \Sigma^* \|_{op} \text{Err}_{f0}(\epsilon) + S_{3,a}^2 \| \Sigma^* \|_{op} \text{Err}_{f0}(\epsilon)^2,
\]

(30)

\[
\| \bar{\Sigma} - \Sigma^* \|_{\max} \leq 4S_{3,a} \| \Sigma^* \|_{\max} \text{Err}_{f0}(\epsilon) + 2S_{3,a}^2 \| \Sigma^* \|_{\max} \text{Err}_{f0}(\epsilon)^2,
\]

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

(ii) Let \( \bar{\theta} = \arg\min_{\theta \in \Theta_0} D_f(P_\theta \| P_\theta) \). Then the statements in (i) hold with \( \text{Err}_{f0}(\epsilon) \) replaced by \( \text{Err}_{f0}(\epsilon) = 2\epsilon \) throughout.

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

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

(31)

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

(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

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

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:

\[
d(\bar{\theta}, \theta^*) \leq TV(P_{\theta}, P_\epsilon) + TV(P_\epsilon, P_{\theta^*}) \leq TV(P_{\theta}, P_\epsilon) + \Delta_1(\epsilon)
\]

\[
\leq \sqrt{2(f''(1))^{-1}D_f(P_\epsilon \| P_{\theta}) + \Delta_1(\epsilon)},
\]

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

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

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

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

where $\text{Err}_{f0}(\epsilon) = \sqrt{-2(f''(1))^{-1}f'(1-\epsilon_0)\epsilon + \epsilon}$. The location result then follows from Proposition 29 provided that $\text{Err}_{f0}(\epsilon) \leq a$ for a constant $a \in [0,1/2)$. The variance matrix result follows if $\text{Err}_{h0}(\epsilon) \leq a/(1 + S_{1,a})$.

(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:

$$d(\bar{\theta}, \theta^*) \leq \text{TV}(P_{\bar{\theta}}, P_\epsilon) + \text{TV}(P_\epsilon, P_{\theta^*}) \leq 2\text{TV}(P_\epsilon, P_{\theta^*}) \leq 2\epsilon.$$ 

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

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

### 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

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

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

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

$$\text{Err}_{n1}(n, p, \delta, \epsilon) = 3\epsilon + 2\sqrt{\epsilon/(n\delta)} + \lambda_{12} + \lambda_{11},$$

where $\lambda_1$ is allowed to depend on $\lambda_{11}$ in the following result.
Proposition 21 Assume that \( \|\Sigma\|_{\max} \leq M_1 \) and \( \epsilon \leq 1/5 \). Let \( \hat{\theta} = (\hat{\mu}, \hat{\Sigma}) \) be a solution to (20) with \( \lambda_1 \geq C_{sp13} M_{11} \lambda_1 \) where \( M_{11} = M_1^{1/2}(M_1^{1/2} + 2\sqrt{2\pi}) \) and \( C_{sp13} = (5/3)(C_{sp11} \vee C_{sp12}) \), depending on universal constants \( C_{sp11} \) and \( C_{sp12} \) in Lemma 30 in Appendix C. If \( \sqrt{\epsilon(1-\epsilon)/(8\delta)} \leq 1/5 \) and \( \text{Err}_{h1}(n, p, \delta, \epsilon) \leq a \) for a constant \( a \in (0, 1/2) \), then the following holds with probability at least \( 1 - 7\delta \) uniformly over contamination distribution \( Q \),

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

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 \( S_{6,a} = S_5(1 + S_{4,a}/2), S_5 = 2\sqrt{2\pi} - 2^{e^{-2/M_1}} - 1 \), and \( S_7 = 4\{\frac{1}{\sqrt{2\pi M_1}}e^{-1/(8M_1)}\} \leq 1 - 2^{e^{-1/(8M_1)}} \) \(-\infty\).

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, \( \lim_{M_1 \to \infty} S_{4,a}/M_1^{1/2} = 2\log(2/\sqrt{(1-2a)})/a \) and \( \lim_{M_1 \to \infty} S_{8,a}/M_1^{5/2} = 8\pi \sqrt{\log(2/\sqrt{(1-2a)})}/a \), 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, \( M_1 \) in \( \text{Err}_{h1}(n, p, \delta, \epsilon) \) 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,

\[
d(\hat{\theta}, \theta^*) - \Delta_{12} \leq \max_{\gamma \in \Gamma} \{ K_{HG}(P_n, P_{\theta}; h_{\gamma,\hat{\theta}}) - \lambda_1 \text{pen}_1(\gamma) \} \leq \Delta_{11}, \tag{37}
\]

where \( \Delta_{11} \) and \( \Delta_{12} \) 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 \( 1 - 5\delta \),

\[
\max_{\gamma \in \Gamma} \{ K_{HG}(P_n, P_{\theta}; h_{\gamma,\hat{\theta}}) - \lambda_1 \text{pen}_1(\gamma) \}
\leq \max_{\gamma \in \Gamma} \{ K_{HG}(P_n, P_{\theta^*}; h_{\gamma,\theta^*}) - \lambda_1 \text{pen}_1(\gamma) \}
\leq \max_{\gamma \in \Gamma} \{ \Delta_{11} + \text{pen}_1(\gamma)\hat{\Delta}_{11} - \lambda_1 \text{pen}_1(\gamma) \}. \tag{38}
\]

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 \( \gamma \in \Gamma \),

\[
K_{HG}(P_n, P_{\theta^*}; h_{\gamma,\theta^*}) \leq \Delta_{11} + \text{pen}_1(\gamma)\hat{\Delta}_{11},
\]
where $\Delta_{11} = 2(\epsilon + \sqrt{\epsilon/(n\delta)})$, $\tilde{\Delta}_{11} = C_{sp13}M_{11}\lambda_{11}$. From (38)–(39), the upper bound in (37) holds with probability at least $1 - 5\delta$, provided that the tuning parameter $\lambda_1$ 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 $1 - 2\delta$,

$$
\max_{\gamma \in \Gamma} \left\{ K_{HG}(P_n, P_\hat{\theta}; h_{\gamma, \hat{\mu}}) - \lambda_1 \text{ pen}_1(\gamma) \right\} \\
\geq \max_{\gamma \in \Gamma_0} \left\{ K_{HG}(P_n, P_\hat{\theta}; h_{\gamma, \hat{\mu}}) - \lambda_1 \text{ pen}_1(\gamma) \right\} \\
\geq \max_{\gamma \in \Gamma_0} \left\{ E_{P_n} h_{\gamma, \hat{\mu}}(x) - E_{P_\hat{\theta}} h_{\gamma, \hat{\mu}}(x) \right\} - \tilde{\Delta}_{12} - \lambda_1. 
$$

(40)

Inequality (40) holds provided that $\Gamma_0$ is a subset of $\Gamma$.

Take $\Gamma_0 = \{ \gamma \in \Gamma_{rp} : \gamma_0 = 0, \text{pen}_1(\gamma) = 1 \}$, where $\Gamma_{rp}$ is the subset of $\Gamma$ 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 $\gamma \in \Gamma_0$,

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

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 $d(\hat{\theta}, \theta^*) = \max_{\gamma \in \Gamma_0} \{ E_{P_n} h_{\gamma, \hat{\mu}}(x) - E_{P_\hat{\theta}} h_{\gamma, \hat{\mu}}(x) \}$.

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

$$
\max_{\gamma \in \Gamma_{rp}, \text{pen}_1(\gamma) = 1} \left\{ E_{P_n} h_{\gamma, \hat{\mu}}(x) - E_{P_\hat{\theta}} h_{\gamma, \hat{\mu}}(x) \right\} \leq \text{Err}_{h_1}(n, p, \delta, \epsilon). 
$$

(42)

where

$$
\text{Err}_{h_1}(n, p, \delta, \epsilon) = 3\epsilon + 2\sqrt{\epsilon/(n\delta)} + \lambda_{12} + \lambda_1.
$$

The desired result then follows from Proposition 42: provided $\text{Err}_{h_1}(n, p, \delta, \epsilon) \leq a$, inequality (42) implies that

$$
\|\hat{\mu} - \mu^*\|_\infty \leq S_4a\text{Err}_{h_1}(n, p, \delta, \epsilon), \quad \|\hat{\Sigma} - \Sigma^*\|_{\max} \leq S_8a\text{Err}_{h_1}(n, p, \delta, \epsilon).
$$

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.

32
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 ($\alpha_d, \alpha_g$) 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

<table>
<thead>
<tr>
<th>Algorithm 3: Penalized logit $f$-GAN or hinge GAN (in detail)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require</strong></td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>2. Learning rates ($\alpha_d, \alpha_g$) for the discriminator and the generator;</td>
</tr>
<tr>
<td>3. Learning rate decay parameters ($d, r$) for the generator;</td>
</tr>
<tr>
<td>4. Numbers of training steps ($s_d, s_g$) for the discriminator and generator;</td>
</tr>
<tr>
<td>5. Base penalty level $\lambda_0$ so that $\lambda_1 = \lambda_0 \sqrt{\log(p)/n}$, $\lambda_2 = \lambda_0 \sqrt{p/n}$, and $\lambda_3 = \lambda_0 \sqrt{p^2/n}$.</td>
</tr>
<tr>
<td>6. Mini-batch size $m$ and number of epochs $T$.</td>
</tr>
<tr>
<td><strong>Initialization</strong></td>
</tr>
<tr>
<td>1. Initialize $\mu_0$ by the median of $X$. Initialize discriminator intercept $\gamma_0$ by 0.01.</td>
</tr>
<tr>
<td>2. Initialize $\Sigma_{1/2}^0$ and the discriminator parameters ($\gamma_1, \gamma_2$) randomly by Xavier uniform (Glorot and Bengio, 2010).</td>
</tr>
<tr>
<td>for $t = 1 \ldots T$ do</td>
</tr>
<tr>
<td>for $u = 1 \ldots T/m$ do</td>
</tr>
<tr>
<td>Draw mini-batch $(x_1, \ldots, x_m)$ from real data without replacement;</td>
</tr>
<tr>
<td>for $s = 1 \ldots s_d$ do</td>
</tr>
<tr>
<td>Generate $(z_1, \ldots, z_m)$ from $N(0, I)$ and the fake data $\mu_{t-1} + \Sigma_{t-1}^{1/2} z_i$, $i = 1, \ldots, m$;</td>
</tr>
<tr>
<td>$g_\gamma \leftarrow \nabla_\gamma K(\theta_{t-1}, \gamma_t; \lambda); \ g_\gamma \leftarrow g_\gamma/|g_\gamma|_2$;</td>
</tr>
<tr>
<td>Update $\gamma_t$ with gradient $g_\gamma$ using the Adam algorithm (Kingma and Ba, 2015) with learning rate $\alpha_d$.</td>
</tr>
<tr>
<td>end</td>
</tr>
<tr>
<td>for $s = 1 \ldots s_g$ do</td>
</tr>
<tr>
<td>Generate $(z_1, \ldots, z_m)$ from $N(0, I)$ and the fake data $\mu_{t-1} + \Sigma_{t-1}^{1/2} z_i$, $i = 1, \ldots, m$;</td>
</tr>
<tr>
<td>$g_\theta \leftarrow \nabla_\theta K(\theta, \gamma_t; \lambda); \ g_\theta \leftarrow g_\theta/|g_\theta|_2$;</td>
</tr>
<tr>
<td>Update $\theta_t$ with gradient $g_\theta$ using the Adam algorithm with learning rate $\alpha_g$.</td>
</tr>
<tr>
<td>end</td>
</tr>
<tr>
<td>Decaying the generator learning rate: $\alpha_g \leftarrow r\alpha_g$ after every $d$ epochs.</td>
</tr>
<tr>
<td>end</td>
</tr>
<tr>
<td>end</td>
</tr>
</tbody>
</table>

33
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 $\tau$ 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

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

to

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

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 $\lambda_0$ 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 $\epsilon$ 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 $\lambda_0$ 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 $\lambda_0$ in the range studied, the relative levels of estimation errors are less sensitive to the choice of $\lambda_0$. Hence we decide to use the same $\lambda_0$ 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$(2.25c, (1/3)l_p)$ or (B) Cauchy$(51p, 5l_p)$. 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 $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. 
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.

### 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 \( \tau \) and Spearman’s \( \rho \)) 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.

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 \( \| \cdot \|_{\text{max}} \).

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 \( \| \cdot \|_F \).
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 $\sqrt{\log(p)/n}$ ($\epsilon$ term aside) has been established for our three $L_1$ penalized methods as well as Kendall’s $\tau$ and Spearman’s $\rho$ (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).
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 $\epsilon$, where outliers do not overlap with the uncontaminated data. The errors from the robust $f$-divergence minimization exhibit square-root dependency on $\epsilon$, whereas those from the TV minimization exhibit linear dependency on $\epsilon$. 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 $\epsilon$. Nevertheless, we also find that despite the worst-case dependency on $\epsilon$, 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 $R_{2,b} = \sup_{|u| \leq b} \frac{d}{du} f'(e^u)$ and $R_{3,b} = R_{31,b} + R_{32,b}$ with $R_{31,b} = \sup_{|u| \leq b} \frac{d^2}{du^2} \{-f'(e^u)\}$.
and $R_{32,b} = \sup_{|u| \leq b} \frac{d^2}{du^2} f^\#(e^u)$. For $\delta \in (0,1)$, define
\[
\lambda_{11} = \sqrt{\frac{2\log(5p) + \log(\delta^{-1})}{n}} + \frac{2\log(5p) + \log(\delta^{-1})}{n},
\]
\[
\lambda_{12} = C_{\text{rad}4}\sqrt{\frac{4\log(2p(p+1))}{n}} + \frac{2\log(\delta^{-1})}{n},
\]
where $C_{\text{rad}4} = C_{sg6}C_{\text{rad}3}$, depending on universal constants $C_{sg6}$ and $C_{\text{rad}3}$ in Lemmas 70 and Corollary 82 in Appendix E. Denote
\[
\text{Err}_{f_1}(n,p,\delta,\epsilon) = (f''(1))^{-1}\left\{ -f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) - f'(e^{-b_1})\sqrt{\epsilon} + \frac{1}{2} R_{3,b_1}(\sqrt{\epsilon} + \sqrt{1/n}) + R_{2,b_1}\lambda_{12} + \lambda_1 \right\},
\]
where $b_1 = \sqrt{\epsilon} + \sqrt{1/n}$. Note that $R_{2,b_1}, R_{3,b_1}$ are bounded provided that $b$ is bounded, because $f$ is three-times continuously differentiable as required in Assumption 2.

**Proposition 23** Assume that $\|\Sigma^*\|_{\text{max}} \leq M_1$, and $f$ satisfies Assumptions 1–2. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (17) with $\lambda_1 \geq C_{sp13} R_{1} M_{11} \lambda_{11}$, where $M_{11} = M_1^{1/2}(M_1^{1/2} + 2\sqrt{2\pi})$ and $C_{sp13} = (5/3)(C_{sp11} \lor C_{sp12})$, 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 $\text{Err}_{f_1}(n,p,\delta,\epsilon) \leq a$ for a constant $a \in (0,1/2)$, then we have that with probability at least $1 - 7\delta$,
\[
\|\hat{\mu} - \mu^*\|_{\infty} \leq S_{4,a} \text{Err}_{f_1}(n,p,\delta,\epsilon),
\]
\[
\|\hat{\Sigma} - \Sigma^*\|_{\text{max}} \leq S_{8,a} \text{Err}_{f_1}(n,p,\delta,\epsilon),
\]
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 $S_{6,a} = S_5(1 + S_{4,a}/2)$, $S_5 = 2\sqrt{2\pi}(1 - e^{-2/M_1})^{-1}$, and $S_7 = 4\{(\frac{1}{\sqrt{2\pi M_1}}e^{-1/(8M_1)}) \lor (1 - 2e^{-1/(8M_1)})\}^{-2}$.

**Proof** [Proof of Proposition 23]

The main strategy of our proof is to show that the following inequalities hold with high probabilities,
\[
d(\hat{\theta}, \theta^*) - \Delta_{12} \leq \max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \text{pen}_1(\gamma) \right\} \leq \Delta_{11},
\] (43)
where $\Delta_{11}$ and $\Delta_{12}$ 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 $\|\hat{\Sigma} - \Sigma^*\|_{\text{max}}$.

(Step 1) For the upper bound in (43), we show that with probability at least $1 - 5\delta$,
\[
\max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - \lambda_1 \text{pen}_1(\gamma) \right\} \leq \max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) - \lambda_1 \text{pen}_1(\gamma) \right\}.
\] (44)

(Step 2) For the lower bound in (43), we show that with probability at least $1 - 2\delta$,
\[
\max_{\gamma \in \Gamma} \left\{ \Delta_{11} + \text{pen}_1(\gamma) \Delta_{11} - \lambda_1 \text{pen}_1(\gamma) \right\}.
\] (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 \( \gamma \in \Gamma \),
\[
K_f(P_n, P_{\theta^*}; h_{\gamma, \hat{\mu}}) \leq \Delta_{11} + \text{pen}_1(\gamma) \tilde{\Delta}_{11},
\]
where \( \Delta_{11} = -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) \), \( \tilde{\Delta}_{11} = C_{sp}3R_1M_{11}\lambda_{11} \), and
\[
\lambda_{11} = \sqrt{\frac{2\log(5)p + \log(\delta^{-1})}{n}} + \frac{2\log(5)p + \log(\delta^{-1})}{n}.
\]
Note that \( \lambda_{11} \) 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 \( \lambda_1 \) is chosen such that \( \lambda_1 \geq \Delta_{11} \).

(Step 2) For the lower bound in (43), we show that with probability at least \( 1 - 2\delta \),
\[
\max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - \lambda_1 \text{pen}_1(\gamma) \right\} \\
\geq \max_{\gamma \in \Gamma_0} \left\{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - \lambda_1 \text{pen}_1(\gamma) \right\} \\
\geq \max_{\gamma \in \Gamma_0} f''(1) \left\{ E_{P_{\theta^*}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta^*}} h_{\gamma, \hat{\mu}}(x) \right\} - \tilde{\Delta}_{12} - \lambda_1 b_1. 
\]
(46)

Inequality (46) holds provided that \( \Gamma_0 \) is a subset of \( \Gamma \). As a subset of the pairwise spline class \( H_{sp} \), define a class of pairwise ramp functions, \( H_{tp} \), such that each function in \( H_{tp} \) can be expressed as, for \( x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p \),
\[
h_{tp, \beta, c}(x) = \beta_0 + \sum_{j=1}^p \beta_{1j} \text{ramp}(x_j - c_j) + \sum_{1 \leq i \neq j \leq p} \beta_{2, ij} \text{ramp}(x_i)\text{ramp}(x_j),
\]
where \( \text{ramp}(t) = \frac{1}{2}(t+1)_+ - \frac{1}{2}(t-1)_+ \) for \( t \in \mathbb{R} \), \( c = (c_1, \ldots, c_p)^T \) with \( c_j \in \{0, 1\} \), and \( \beta = (\beta_0, \beta_1^T, \beta_2^T)^T \) with \( \beta_1 = (\beta_{1j} : j = 1, \ldots, p)^T \) and \( \beta_2 = (\beta_{2, ij} : 1 \leq i \neq j \leq p)^T \). For symmetry as in \( \gamma_2 \), assume that the coefficients in \( \beta_2 \) are symmetric, \( \beta_{2, ij} = \beta_{2, ji} \) for any \( i \neq j \). By the definition of \( \text{ramp}(\cdot) \), each function \( h_{tp, \beta, c}(x) \) can be represented as \( h_{\gamma}(x) \) in the spline class \( H_{sp} \), where \( \beta \) and \( \gamma \) 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 \( \gamma_2 \) or in \( \beta_2 \). Denote as \( \Gamma_{tp} \) the subset of \( \Gamma \) such that \( H_{tp} = \{ h_{\gamma}(x) : \gamma \in \Gamma_{tp} \} \).

Take \( \Gamma_0 = \{ \gamma \in \Gamma_{tp} : \gamma_0 = 0, \text{pen}_1(\gamma) = b_1 \} \) for some fixed \( b_1 > 0 \). Inequality (47) follows from Proposition 37: it holds with probability at least \( 1 - 2\delta \) that for any \( \gamma \in \Gamma_0 \),
\[
K_f(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) \geq f''(1) \left\{ E_{P_{\theta^*}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta^*}} h_{\gamma, \hat{\mu}}(x) \right\} - \tilde{\Delta}_{12},
\]
where \( \tilde{\Delta}_{12} = -f'(e^{-b_1})\epsilon + \frac{1}{2}b_1^2R_{3, b_1} + b_1R_{2, b_1}\lambda_{12} \), and
\[
\lambda_{12} = C_{\text{rad}}\sqrt{\frac{4\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}.
\]
Note that \( \lambda_{12} \) 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 \( \Delta_{12} = \Delta_{12} + \lambda_1 b_1 \) and \( d(\hat{\theta}, \theta^*) = f''(1) \max_{\gamma \in \Gamma_0} \{ E_{P_{\theta^*}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta}} h_{\gamma, \hat{\mu}}(x) \} \).

(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 \( \theta^* \). First, due to the linearity of \( h_{\gamma, \hat{\mu}} \) in \( \gamma \), combining the lower and upper bounds in (43) shows that with probability at least \( 1 - 7\delta \),

\[
f''(1)b_1 \max_{\gamma \in \Gamma_{p,pen}(\gamma) = 1} \{ E_{P_{\theta^*}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta}} h_{\gamma, \hat{\mu}}(x) \} \leq -f'(3/5)\left( \epsilon + \sqrt{\epsilon/(n\delta)} \right) - f'(e^{-b_1})\epsilon + \frac{1}{2} b_1^2 R_{3,b_1} + b_1 R_{2,b_1} \lambda_{12} + \lambda_1 b_1.
\]

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

\[
\max_{\gamma \in \Gamma_{p,pen}(\gamma) = 1} \{ E_{P_{\theta^*}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta}} h_{\gamma, \hat{\mu}}(x) \} \leq \text{Err}_f(n, p, \delta, \epsilon),
\]

where

\[
\text{Err}_f(n, p, \delta, \epsilon) = (f''(1))^{-1} \left\{ -f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) - f'(e^{-b_1})\sqrt{\epsilon} + \frac{1}{2} R_{3,b_1}(\sqrt{\epsilon} + 1/\sqrt{n}) + R_{2,b_1} \lambda_{12} + \lambda_1 \right\}.
\]

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

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

\[\blacksquare\]

### B.2 Proof of Theorem 12

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

\[
\lambda_{21} = \sqrt{\frac{5p + \log(\delta^{-1})}{n}}, \quad \lambda_{22} = C_{\text{rad}5} \sqrt{\frac{16p}{n}} + \sqrt{\frac{2p \log(\delta^{-1})}{n}}, \\
\lambda_{31} = \lambda_{21} + \frac{5p + \log(\delta^{-1})}{n}, \quad \lambda_{32} = C_{\text{rad}5} \sqrt{\frac{6(p-1)}{n}} + \sqrt{\frac{(p-1) \log(\delta^{-1})}{n}}.
\]

where \( C_{\text{rad}5} = C_{sg,12} C_{\text{rad}3} \), depending on universal constants \( C_{sg,12} \) and \( C_{\text{rad}3} \) in Lemmas 67 and Corollary 82 in Appendix E. Denote

\[
\text{Err}_f(n, p, \delta, \epsilon) = (\sqrt{2} R_{4,b_2})^{-1} \left\{ -f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) - f'(e^{-b_2})\sqrt{\epsilon} \\
+ 4C_{sg,12} M_2 R_{3,b_2}(\sqrt{\epsilon} + \sqrt{1/(np)}) + R_{2,b_2} \lambda_{22} + \lambda_2 \right\},
\]

\[
\text{Err}_f(n, p, \delta, \epsilon) = (2 R_{4,2b_2})^{-1} \left\{ -f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) - f'(e^{-2b_2})\sqrt{\epsilon} \\
+ (80 C_{sg,12}^2 M_2) R_{3,2b_2}(\sqrt{\epsilon} + \sqrt{1/(np)}) + R_{2,b_2} \lambda_{32} + \lambda_3 / \sqrt{p} \right\},
\]

44
where \( b_2 = \sqrt{\epsilon + 1/(np)} \), \( b_2^\dagger = b_2 \sqrt{2p} \), \( b_3 = \sqrt{\epsilon/p + 1/(np^2)} \), 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

\[
R_{4,b} = \inf_{|u| \leq b} \frac{d}{du} f'(e^u) = \inf_{|u| \leq b} \frac{d}{du} \{ -f'(f'(e^u)) \}
\]

is bounded away from zero provided that \( b \) is bounded.

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

\[
\lambda_2 \geq (5/3)C_{sp21} M_{21}^{1/2} R_1 \lambda_{31}, \quad \lambda_3/\sqrt{p} \geq (25\sqrt{5}/3)C_{sp22} M_1 R_1 \lambda_{31},
\]

where \( M_{21} = M_2^{1/2}(M_2^{1/2} + 2\sqrt{\pi}) \), \( C_{sp21} = \sqrt{2}C_{sg7}C_{sg5} \), and \( C_{sp22} = \sqrt{2\pi}C_{sp21} + C_{sg8} \), depending on universal constants \( C_{sg5}, C_{sg7}, \) and \( C_{sg8} \) in Lemmas 69, 71, and 72 in Appendix E. If \( \epsilon \leq 1/5 \), \( \sqrt{\epsilon f(1-\epsilon)/(n\delta)} \leq 1/5 \), and \( Err_f(\hat{\Sigma}, \hat{\mu}, \Sigma^*, \epsilon) \leq a \) for a constant \( a \in (0, 1/2) \), then we have that with probability at least \( 1 - 8\delta \),

\[
\|\hat{\mu} - \mu^*\|_2 \leq S_{4,a} Err_f(n, p, \delta, \epsilon),
\]

\[
p^{-1/2}\|\hat{\Sigma} - \Sigma^*\|_F \leq S_{9,a} Err_f(n, p, \delta, \epsilon) + S_7 Err_f(n, p, \delta, \epsilon),
\]

where \( S_{4,a} = 2M_1^{1/2}S_{6,a} + \sqrt{2}S_7(S_{4,a} + S_{6,a}) \) 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 \( \lim_{M_2 \to \infty} S_{9,a}/M_2^{5/2} = 8\pi \sqrt{\log(2/(1-2a))}/a \), 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, \( \lambda_2 \in Err_f(n, p, \delta, \epsilon) \) can be set to linearly depend on \( M_2^{1/2} \), and \( \lambda_3 \in Err_f(n, p, \delta, \epsilon) \) 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** The main strategy of our proof is to show that the following inequalities hold with high probabilities,

\[
d(\hat{\theta}, \theta^*) - \Delta_22 \leq \max_{\gamma \in \Gamma} \{ K_f(P_0, P_0; h_{\gamma, \hat{\mu}}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \} \leq \Delta_{21}, \tag{49}
\]

where \( \Delta_{21} \) and \( \Delta_{22} \) 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 \( \lambda_2 \) and \( \lambda_3 \) associated with \( \gamma_1 \) and \( \gamma_2 \) 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 \( \sigma^* \) is the vector of standard deviations from \( \hat{\Sigma} \) or \( \Sigma^* \) respectively. In Steps 4 and 5, we derive the second version of the lower bound in (49) and then deduce an upper bound on \( \|\Sigma - \Sigma^*\|_F \).
(Step 1) For the upper bound in (49), we show that with probability at least $1 - 4\delta$,

$$
\max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\leq \max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\theta'}; h_{\gamma, \mu'}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\leq \max_{\gamma \in \Gamma} \left\{ \Delta_{21} + \text{pen}_2(\gamma_1)\Delta_{21} + \text{pen}_2(\gamma_2)\Delta_{31} - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}.
$$

(50)

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 $\gamma \in \Gamma_1$,

$$
K_f(P_n, P_{\theta'}; h_{\gamma, \mu'}) \leq \Delta_{21} + \text{pen}_2(\gamma_1)\Delta_{21} + \text{pen}_2(\gamma_2)\sqrt{p}\Delta_{31},
$$

where

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

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 $\lambda_2$ and $\lambda_3$ are chosen such that $\lambda_2 \geq \Delta_{21}$ and $\lambda_3 \geq \sqrt{p}\Delta_{31}$.

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

$$
\max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\theta'}; h_{\gamma, \mu'}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\geq \max_{\gamma \in \Gamma_1} \left\{ K_f(P_n, P_{\theta'}; h_{\gamma, \mu'}) - \lambda_2 \text{pen}_2(\gamma_1) \right\}
\geq \max_{\gamma \in \Gamma_{10}} \left\{ K_f(P_n, P_{\theta'}; h_{\gamma, \mu'}) - \lambda_2 \text{pen}_2(\gamma_1) \right\}
\geq \max_{\gamma \in \Gamma_{10}} R_{4, b_2} \left\{ E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) \right\} - \Delta_{22} - \lambda_2 b_2,
$$

(52)

(53)

(54)

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

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

where $\text{ramp}(t) = \frac{1}{2}(t + 1)_+ - \frac{1}{2}(t - 1)_+$ for $t \in \mathbb{R}$, $c = (c_1, \ldots, c_p)^T$ with $c_j \in \{0, 1\}$, and $\beta = (\beta_0, \beta_1^T)^T$ with $\beta_1 = (\beta_{11}, \ldots, \beta_{1p})^T$. Only the main-effect ramp functions are included, while the interaction ramp functions are excluded, in $h_{rp, \beta, c}(x)$. By the definition of $\text{ramp}(\cdot)$, each function $h_{rp, \beta, c}(x)$ can be represented as $h_{\gamma}(x) \in \mathcal{H}_{sp}$ with $\gamma = (\gamma_0, \gamma_1^T)^T \in \Gamma_{rp}$.
such that \( \beta \) and \( \gamma \) satisfy \( \beta_0 = \gamma_0 \) and \( \|\beta_1\|_2 = \sqrt{2}\|\gamma_1\|_2 \). For example, for \( \text{ramp}(x) \), the associated norms are \( \|\beta_1\|_2 = 1 \) and \( \|\gamma_1\|_2 = \sqrt{1/2} \). Denote as \( \Gamma_{\text{rp}1} \) the subset of \( \Gamma_1 \) such that \( \mathcal{H}_{\text{rp}1} = \{ h_\gamma(x) : \gamma \in \Gamma_{\text{rp}1} \} \).

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

\[
K_f(P_n, P_{\hat{\theta}}, h_{\gamma, \hat{\mu}}) \geq R_{4,b_2^1} \left\{ E_{P_\theta} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \right\} - \tilde{\Delta}_{22},
\]

where \( b_2^1 = b_2 \sqrt{2p} \), \( \tilde{\Delta}_{22} = -f'(e^{-b_2^1})\epsilon + 4C_{s_0,12}^2 M_2 b_2^2 R_{3,b_2^1} + b_2 R_{2,b_2^1} \lambda_{22} \), and

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

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

(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 \( 1 - 6\delta \),

\[
R_{4,b_2^1} \max_{\gamma \in \Gamma_{\text{rp}1, \text{pen}_2(\gamma)} = 1} \left\{ E_{P_\theta} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \right\} \leq -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) - f'(e^{-b_2^1})\epsilon + 4C_{s_0,12}^2 M_2 b_2^2 R_{3,b_2^1} + b_2 R_{2,b_2^1} \lambda_{22} + \lambda_2 b_2.
\]

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

\[
\max_{\gamma \in \Gamma_{\text{rp}1, \text{pen}_2(\gamma)} = 1} \left\{ E_{P_\theta} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \right\} \leq \text{Err}_{f_2}(n, p, \delta, \epsilon),
\]

where

\[
\text{Err}_{f_2}(n, p, \delta, \epsilon) = (\sqrt{2} R_{4,b_2^1})^{-1} \left\{ -f'(3/5)(\sqrt{\epsilon + \sqrt{1/(n\delta)})} \right. \\
\left. - f'(e^{-b_2^1})\sqrt{\epsilon + 4C_{s_0,12}^2 M_2 R_{3,b_2^1} (\sqrt{\epsilon + \sqrt{1/(np)})}} + R_{2,b_2^1} \lambda_{22} + \lambda_2 \right\}.
\]

The error bounds for \( (\hat{\mu}, \hat{\sigma}) \) then follows from Proposition 48: provided \( \text{Err}_{f_2}(n, p, \delta, \epsilon) \leq a \), inequality (55) implies that

\[
\|\hat{\mu} - \mu^*\|_2 \leq S_{4,a} \text{Err}_{f_2}(n, p, \delta, \epsilon),
\]

\[
\|\hat{\sigma} - \sigma^*\|_2 \leq S_{6,a} \text{Err}_{f_2}(n, p, \delta, \epsilon).
\]
(Step 4) For the second version of the lower bound in (49), we show that with probability at least $1 - 2\delta$,

$$
\max_{\gamma \in \Gamma} \left\{ K_f(P_n, P_{\hat{\gamma}}; h_{\gamma, \hat{\mu}}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\geq \max_{\gamma \in \Gamma_2} \left\{ K_f(P_n, P_{\hat{\gamma}}; h_{\gamma, \hat{\mu}}) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\geq \max_{\gamma \in \Gamma_2} \left\{ K_f(P_n, P_{\hat{\gamma}}; h_{\gamma, \hat{\mu}}) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\geq \max_{\gamma \in \Gamma_2} f''(1) \left\{ E_{P_{\hat{\gamma}}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\gamma}}} h_{\gamma, \hat{\mu}}(x) \right\} - \tilde{\Delta}_{32} - \lambda_3 b_3,
$$

(58)

(59)

(60)

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

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

where $\text{ramp}(t) = \frac{1}{2}(t + 1)_+ - \frac{1}{2}(t - 1)_+$ for $t \in \mathbb{R}$, and $\beta = (\beta_0, \beta^T_2)$ with $\beta_2 = (\beta_{2ij} : 1 \leq i \neq j \leq p)^T$. In contrast with the function $h_{rp1, \beta, c}(x)$ in $\mathcal{H}_{sp1}$, only the interaction ramp functions are included, while the main-effect ramp functions are excluded, in $h_{rp2, \beta}$. For symmetry as in $\gamma_2$, assume that the coefficients in $\beta_2$ are symmetric, $\beta_{2ij} = \beta_{2ji}$ for any $i \neq j$. By the definition of $\text{ramp}(\cdot)$, each function $h_{rp2, \beta}(x)$ can be represented as $h_{x_2}(x) \in \mathcal{H}_{sp2}$ with $\gamma = (\gamma_0, \gamma_2^T)^T \in \Gamma_{rp2}$, such that $\beta$ and $\gamma$ satisfy $\beta_0 = \gamma_0$ and $\|\beta_2\|_2 = 2\|\gamma_2\|_2$. For example, for $\text{ramp}(x_1) \text{ramp}(x_2)$, the associated norms are $\|\beta_2\|_2 = 1$ and $\|\gamma_2\|_2 = 1/2$. Denote as $\Gamma_{rp2}$ the subset of $\Gamma_2$ such that $h_{rp2} \in \mathcal{H}_{rp2}$.

Take $\Gamma_2 = \{ \gamma \in \Gamma_{rp2} : \text{pen}_2(\gamma) = b_3, E_{P_{\hat{\gamma}}} h_{\gamma, \hat{\mu}}(x) = 0, E_{P_{\hat{\gamma}}} h_{\gamma, \hat{\mu}}(x) \leq 0 \}$ for some fixed $b_3 > 0$. Inequality (60) follows from Proposition 49: it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_2$,

$$
K_f(P_n, P_{\hat{\gamma}}; h_{\gamma, \hat{\mu}}) \geq R_{4, 2b_3} \left\{ E_{P_{\hat{\gamma}}} h(x) - E_{P_{\hat{\gamma}}} h(x) \right\} - \tilde{\Delta}_{32},
$$

where $b_3 = b_3 \sqrt{p(p - 1)}$, $\tilde{\Delta}_{32} = -f''(e^{-2b_3}) + (80C_{sp1, 12}^2 M_2) b_3^2 R_{3, 2b_3} + \sqrt{b_3} R_{2, b_3} \lambda_{32}$, and

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

From (59)–(60), the lower bound in (49) holds with probability at least $1 - 2\delta$, where $\Delta_{22} = \Delta_{32} + \lambda_3 b_3$ and $d(\theta, \theta^*) = R_{4, 2b_3} \max_{\gamma \in \Gamma_2} \left\{ E_{P_{\hat{\gamma}}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\gamma}}} h_{\gamma, \hat{\mu}}(x) \right\}$.

(Step 5) We deduce an upper bound on $\|\Sigma - \Sigma^*\|_F$, by choosing appropriate $b_3$ and relating the moment matching term $d(\theta, \theta^*)$ 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 $1 - 6\delta$,

$$
R_{4, 2b_3}^{1/3} \max_{\gamma \in \Gamma_{rp2}, \text{pen}_2(\gamma) = 1} \left\{ E_{P_{\hat{\gamma}}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\gamma}}} h_{\gamma, \hat{\mu}}(x) \right\}
\leq -f''(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) - f''(e^{-2b_3}) + (80C_{sp1, 12}^2 M_2) b_3^2 R_{3, 2b_3} + \sqrt{b_3} R_{2, b_3} \lambda_{32} + \lambda_3 b_3.
$$
Taking $b_3 = \sqrt{\epsilon/p + 1/(np^2)}$ in the preceding display and rearranging yields

$$\max_{\gamma \in \Gamma_{np^2 \text{-pen}}(\gamma) = 1/2} \left\{ E_{P_{\theta^*}} h_{\gamma,\hat{\mu}}(x) - E_{P_{\theta}} h_{\gamma,\hat{\mu}}(x) \right\} \leq \sqrt{p} \text{Err}_f^3(n, p, \delta, \epsilon), \quad (61)$$

where

$$\text{Err}_f^3(n, p, \delta, \epsilon) = (2R_{4,2b_3}^{-1}) \left\{ - f'(3/5)(\sqrt{\epsilon} + \sqrt{1/(n\delta)}) - f'((e^{-2b_3})/\sqrt{\epsilon} + (80C_{sp2,12}^2 M_2) R_{3,2b_3} (\sqrt{\epsilon} + \sqrt{1/(np)}) + R_{2,b_3} \lambda_{32} + \lambda_3 / \sqrt{p} \right\}.$$

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

$$\frac{1}{\sqrt{p}} \|\hat{\Sigma} - \Sigma^*\|_F \leq 2M_2^{1/2} \|\hat{\sigma} - \sigma^*\|_2 + S_T \left\{ \sqrt{2}\Delta_{\hat{\mu},\hat{\sigma}} + \text{Err}_f^3(n, p, \delta, \epsilon) \right\}$$

$$\leq S_{9,a} \text{Err}_f^2(n, p, \delta, \epsilon) + S_T \text{Err}_f^3(n, p, \delta, \epsilon),$$

where $\Delta_{\hat{\mu},\hat{\sigma}} = (\|\hat{\mu} - \mu^*\|^2_2 + \|\hat{\sigma} - \sigma^*\|^2_2)^{1/2}$ and $S_{9,a} = 2M_2^{1/2} S_{6,a} + \sqrt{2} S_T (S_{4,a} + S_{6,a}).$ \hfill \[1.5pt]\[0.5pt]

### 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

$$\text{Err}_{h2}(n, p, \delta, \epsilon) = 3\epsilon (2p)^{1/2} + 2\sqrt{2p^2c/(n\delta)} + \lambda_2 + \lambda_{22},$$

$$\text{Err}_{h3}(n, p, \delta, \epsilon) = 3\epsilon \sqrt{p - 1} + 2\sqrt{\epsilon (p - 1)/(n\delta)} + \lambda_{32}/2 + (25\sqrt{5}/6) C_{sp22} M_2 \lambda_{31}.$$

**Proposition 26** Assume that $\|\Sigma^*\|_{\text{op}} \leq M_2$. Let $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ be a solution to (21) with $\lambda_3 / \sqrt{p} \geq (25\sqrt{5}/3) C_{sp22} M_2 \lambda_{31}$ and $\lambda_2 \geq (5/3) C_{sp21} M_2^{1/2} \lambda_{21}$, where $M_2$, $C_{sp21}$, and $C_{sp22}$ are defined as in Proposition 24. If $\epsilon \leq 1/5$, $\sqrt{\epsilon (1 - \epsilon)/(n\delta)} \leq 1/5$ and $\text{Err}_{h2}(n, p, \delta, \epsilon) \leq a$ for a constant $a \in (0, 1/2)$, then we have that with probability at least $1 - 8\delta$,

$$\|\hat{\mu} - \mu^*\|^2_2 \leq S_{4,a} \text{Err}_{h2}(n, p, \delta, \epsilon),$$

$$p^{-1/2} \|\hat{\Sigma} - \Sigma^*\|_F \leq S_{9,a} \text{Err}_{h2}(n, p, \delta, \epsilon) + S_T \text{Err}_{h3}(n, p, \delta, \epsilon),$$

where $(S_{4,a}, S_{6,a}, S_T, S_{9,a})$ 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,

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

where $\Delta_{21}$ and $\Delta_{22}$ 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 − 4δ,
\[
\max_{\gamma \in \Gamma} \left\{ K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\leq \max_{\gamma \in \Gamma} \left\{ K_{HG}(P_n, P_{\theta^*}; h_{\gamma, \mu^*}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\leq \max_{\gamma \in \Gamma} \left\{ \Delta_{21} + \text{pen}_2(\gamma_1)\hat{\Delta}_{21} + \text{pen}_2(\gamma_2)\hat{\Delta}_{31} - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}.
\]  
(63) 
Inequality (63) follows from the definition of \( \hat{\theta} \). Inequality (64) follows from Proposition 54: it holds with probability at least 1 − 4δ that for any \( \gamma \in \Gamma_1 \),
\[
K_{HG}(P_n, P_{\theta^*}; h_{\gamma, \mu^*}) \leq \Delta_{21} + \text{pen}_2(\gamma_1)\hat{\Delta}_{21} + \text{pen}_2(\gamma_2)\sqrt{\hat{\Delta}_{31}},
\]
where
\[
\Delta_{21} = 2(\epsilon + \sqrt{\epsilon/(n\delta)}), \quad \hat{\Delta}_{21} = (5/3)C_{sp21}M_{21}^{1/2}\lambda_{21}, \quad \hat{\Delta}_{31} = (25\sqrt{5}/3)C_{sp22}M_{21}\lambda_{31},
\]
and \( \lambda_{21} \) and \( \lambda_{31} \) are the same as in the proof of Theorem 12. Note that \( \hat{\Delta}_{21} \) and \( \hat{\Delta}_{31} \) 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 − 4δ, provided that the tuning parameters \( \lambda_2 \) and \( \lambda_3 \) are chosen such that \( \lambda_2 \geq \hat{\Delta}_{21} \) and \( \lambda_3 \geq \sqrt{\hat{\Delta}_{31}} \).

(Step 2) For the first version of the lower bound in (62), we show that with probability at least 1 − 2δ,
\[
\max_{\gamma \in \Gamma} \left\{ K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\}
\geq \max_{\gamma \in \Gamma_1} \left\{ K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - \lambda_2 \text{pen}_2(\gamma_1) \right\}
\geq \max_{\gamma \in \Gamma_{10}} \left\{ K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - \lambda_2 (2p)^{-1/2} \right\}
\geq \max_{\gamma \in \Gamma_{10}} \left\{ \mathbb{E}_{P_{\theta^*}, h_{\gamma, \hat{\mu}(x)}} - \mathbb{E}_{P_{\hat{\theta}} h_{\gamma, \hat{\mu}(x)}} - \hat{\Delta}_{22} - \lambda_2 (2p)^{-1/2} \right\},
\]  
(65) 
(66) 
(67) 
where \( \Gamma_1 = \{(\gamma_0, \gamma_1^T, \gamma_2) : \gamma_2 = 0\} \). Inequality (65) follows because \( \Gamma_1 \) is defined as a subset of \( \Gamma \) such that \( \gamma_2 = 0 \) and hence \( \text{pen}_2(\gamma_2) = 0 \) for \( \gamma \in \Gamma_1 \).

Take \( \Gamma_{10} = \{ \gamma \in \Gamma_{\text{rp}} : \gamma_0 = 0, \text{pen}_2(\gamma) = (2p)^{-1/2} \} \), where \( \Gamma_{\text{rp}} \) is the subset of \( \Gamma \) 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 − 2δ that for any \( \gamma \in \Gamma_{10} \),
\[
K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) \geq \mathbb{E}_{P_{\theta^*}, h_{\gamma, \hat{\mu}(x)}} - \mathbb{E}_{P_{\hat{\theta}} h_{\gamma, \hat{\mu}(x)}} - \hat{\Delta}_{22},
\]
where \( \hat{\Delta}_{22} = \epsilon + \lambda_{22}(2p)^{-1/2} \), and
\[
\lambda_{22} = C_{\text{rad}} \sqrt{\frac{16p}{n} + \sqrt{\frac{2p\log(\delta^{-1})}{n}}}. 
\]
Note that \( \lambda_{22} \) is the same as in the proof of Theorem 12. From (65)–(67), the lower bound in (62) holds with probability at least 1 − 2δ, where \( \Delta_{22} = \hat{\Delta}_{22} + \lambda_{22}(2p)^{-1/2} \) and 
\[
d(\hat{\theta}, \theta^*) = \mathbb{E}_{P_{\theta^*}, h_{\gamma, \hat{\mu}(x)}} - \mathbb{E}_{P_{\hat{\theta}} h_{\gamma, \hat{\mu}(x)}}.
\]
(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 \( 1 - 6\delta \),

\[
(2p)^{-1/2} \max_{\gamma \in \Gamma_{\text{pen}}(\gamma) = \sqrt{1/2}} \left\{ E_{P_\theta} h_{\gamma, \hat{\mu}}(x) - E_{P_\theta} h_{\gamma, \hat{\mu}}(x) \right\} \\
\leq 3\epsilon + 2\sqrt{\epsilon / (n\delta)} + (\lambda_2 + \lambda_{22})(2p)^{-1/2},
\]

which gives

\[
\max_{\gamma \in \Gamma_{\text{pen}}(\gamma) = \sqrt{1/2}} \left\{ E_{P_\theta} h_{\gamma, \hat{\mu}}(x) - E_{P_\theta} h_{\gamma, \hat{\mu}}(x) \right\} \leq \text{Err}_{h2}(n, p, \delta, \epsilon), \tag{68}
\]

where

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

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

\[
\|\hat{\mu} - \mu^*\|_2 \leq S_{4,a} \text{Err}_{h2}(n, p, \delta, \epsilon), \\
\|\hat{\sigma} - \sigma^*\|_2 \leq S_{6,a} \text{Err}_{h2}(n, p, \delta, \epsilon).
\]

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

\[
\max_{\gamma \in \Gamma} \left\{ K_{\text{HG}}(P_n, P_\theta; h_{\gamma, \hat{\mu}}) - \lambda_2 \text{pen}_2(\gamma_1) - \lambda_3 \text{pen}_2(\gamma_2) \right\} \\
\geq \max_{\gamma \in \Gamma_2} \left\{ K_{\text{HG}}(P_n, P_\theta; h_{\gamma, \hat{\mu}}) - \lambda_3 \text{pen}_2(\gamma_2) \right\} \tag{69} \\
\geq \max_{\gamma \in \Gamma_{20}} \left\{ K_{\text{HG}}(P_n, P_\theta; h_{\gamma, \hat{\mu}}) \right\} - \lambda_3 (4q)^{-1/2} \tag{70} \\
\geq \max_{\gamma \in \Gamma_{20}} \left\{ E_{P_\theta} h_{\gamma, \hat{\mu}}(x) - E_{P_\theta} h_{\gamma, \hat{\mu}}(x) \right\} - \tilde{\Delta}_{32} - \lambda_3 (4q)^{-1/2}, \tag{71}
\]

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

Take \( \Gamma_{20} = \{ \gamma \in \Gamma_{\text{pen}} : \gamma_0 = 0, \text{pen}_2(\gamma) = (4q)^{-1/2} \} \) for \( q = p(1 - p) \), where \( \Gamma_{\text{pen}} \) is the subset of \( \Gamma_2 \) 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 \( \gamma \in \Gamma_{20} \),

\[
K_{\text{HG}}(P_n, P_\theta; h_{\gamma, \hat{\mu}}) \geq E_{P_\theta} h(x) - E_{P_\theta} h(x) - \tilde{\Delta}_{32},
\]

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

\[
\lambda_{32} = C_{\text{rad4}} \sqrt{\frac{6(p - 1)}{n}} + \sqrt{\frac{(p - 1) \log(\delta^{-1})}{n}}.
\]
Note that $\lambda_{32}$ 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 $d(\hat{\theta}, \theta^*) = \max_{\gamma \in \Gamma} \{E_{P_{\theta}} h_{\gamma,\hat{\mu}}(x) - E_{P_{\theta}} h_{\gamma,\hat{\mu}}(x)\}$.

(Step 5) We deduce an upper bound on $\|\hat{\Sigma} - \Sigma^*\|_F$, 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 $1 - 6\delta$,

$$
(4q)^{-1/2} \max_{\gamma \in \Gamma_{p2,\text{pen}_2}(\gamma)=1} \left\{ E_{P_{\theta}} h_{\gamma,\hat{\mu}}(x) - E_{P_{\theta}} h_{\gamma,\hat{\mu}}(x) \right\} \leq 3\epsilon + 2\sqrt{\epsilon/(n\delta)} - \sqrt{p} \lambda_{32}(4q)^{-1/2} - (25/3)\sqrt{5}pC_{sp22}M_{21}\lambda_{31}(4q)^{-1/2},
$$

which gives

$$
\max_{\gamma \in \Gamma_{p2,\text{pen}_2}(\gamma)=1/2} \left\{ E_{P_{\theta}} h_{\gamma,\hat{\mu}}(x) - E_{P_{\theta}} h_{\gamma,\hat{\mu}}(x) \right\} \leq \sqrt{p} \text{Err}_{h3}(n, p, \delta, \epsilon), \tag{72}
$$

where

$$
\text{Err}_{h3}(n, p, \delta, \epsilon) = 3\epsilon \sqrt{p - 1} + 2\sqrt{\epsilon(p - 1)/(n\delta)} + \lambda_{32}/2 - (25/3)\sqrt{5}pC_{sp22}M_{21}\lambda_{31}
$$

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

$$
\frac{1}{\sqrt{p}} \|\hat{\Sigma} - \Sigma^*\|_F \leq 2M_{21}^{1/2}\|\hat{\sigma} - \sigma^*\|_2 + S_{\gamma}(\sqrt{2}\Delta_{\mu,\hat{\sigma}} + \text{Err}_{h3}(n, p, \delta, \epsilon))
$$

$$
\leq S_{9,a}\text{Err}_{h2}(n, p, \delta, \epsilon) + S_{7}\text{Err}_{h3}(n, p, \delta, \epsilon),
$$

where $\Delta_{\mu,\hat{\sigma}} = (\|\hat{\mu} - \mu^*\|_2^2 + \|\hat{\sigma} - \sigma^*\|_2^2)^{1/2}$ and $S_{9,a} = 2M_{21}^{1/2}S_{6,a} + \sqrt{2}S_{7}(S_{4,a} + S_{6,a})$.

\[\square\]

\textbf{B.4 Proof of Corollary 18}

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

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

where $\text{pen}(\gamma; \lambda)$ 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 $\theta$ is the minimizer of

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

which is a function of $\theta$ 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 $\Gamma$. 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 $1 - \delta$,

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

$$\leq \max_{\gamma \in \Gamma} \{ -f'(1 - \hat{\epsilon}) + R_1 \left| E_{P_{\hat{\mu}, n}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\mu}} h_{\gamma, \hat{\mu}}(x)} - \text{pen}(\gamma; \lambda) \right| \}$$

$$\leq \Delta_1 + \max_{\gamma \in \Gamma} \{ R_1 \left| E_{P_{\hat{\mu}, n}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\mu}} h_{\gamma, \hat{\mu}}(x)} \right| - \text{pen}(\gamma; \lambda) \},$$

(74)

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 $\Delta_{11}$ and $\Delta_{21}$ in the proofs of Theorems 11 and 12.

Similarly as in Proposition 33 or 44, the term $|E_{P_{\hat{\mu}, n}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\mu}} h_{\gamma, \hat{\mu}}(x)}|$ can be controlled in terms of the $L_1$ or $L_2$ norms of $(\gamma_1, \gamma_2)$, using Lemma 30 or 43. Then for $\text{pen}(\gamma; \lambda)$ 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 $1 - 6\delta$,

$$R_1 \left| E_{P_{\hat{\mu}, n}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\mu}} h_{\gamma, \hat{\mu}}(x)} \right| - \text{pen}(\gamma; \lambda) \leq 0.$$

(76)

provided that the tuning parameters $\lambda_1$ or $(\lambda_2, \lambda_3)$ 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

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

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

**Proof** Because $x^2$ is convex in $x$, by Jensen’s inequality, we have

$$\text{TV}(P, Q)^2 \leq \int f_{TV}^2(p/q) dQ,$$

where $f_{TV}(t) = (1 - t)_+$ and $p/q$ is the density ratio $dP/dQ$. Note that $D_f(P||Q)$ can be equivalently obtained as $D_f(P||Q)$, where $\tilde{f}(t) = f(t) - f'(1)(t - 1)$. Therefore, it suffices to show that $\tilde{f}(t) \geq \frac{C_f}{2} f_{TV}^2(t)$ for $t \in (0, \infty)$. 53
By a Taylor expansion of $f$, we have

$$f(t) = f(1) + f'(1)(t - 1) + \frac{f''(\bar{t})}{2}(t - 1)^2$$

$$\geq f'(1)(t - 1) + \frac{C_f}{2}(t - 1)^2,$$  \hspace{1cm} (77)

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

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

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

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

Combining the two cases completes the proof. \hfill \blacksquare

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

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

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

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

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

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

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

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

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

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

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

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

where $S_{1,a} = \{ \Phi'(\Phi^{-1}(1/2 + a)) \}^{-1}$ as in Lemma 28.
(ii) If further \( d(\bar{\theta}, \theta^*) \leq a/(1 + S_{1,a}) \), then

\[
\|\bar{\Sigma} - \Sigma^*\|_{op} \leq 2S_{3,a}\|\Sigma^*\|_{op}d(\bar{\theta}, \theta^*) + S_{3,a}^2\|\Sigma^*\|_{op}(d(\bar{\theta}, \theta^*))^2,
\]

\[
\|\bar{\Sigma} - \Sigma^*\|_{\max} \leq 4S_{3,a}\|\Sigma^*\|_{\max}d(\bar{\theta}, \theta^*) + 2S_{3,a}^2\|\Sigma^*\|_{\max}(d(\bar{\theta}, \theta^*))^2,
\]

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

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

\[
TV(P_1, P_2) = \sup_{A \in A} |P_1(A) - P_2(A)|,
\]

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 \( |P_{\theta}(A) - P_{\theta^*}(A)| \).

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

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

\[
P_{\mu^*, \Sigma^*}(u^T X \leq u^T \bar{\mu}) - P_{\hat{\mu}, \bar{\Sigma}}(u^T X \leq u^T \bar{\mu}) \leq d(\bar{\theta}, \theta^*).
\]

For nonzero \( u \in \mathbb{R}^p \), because \( u^T \bar{\Sigma}u \neq 0 \) and \( u^T \Sigma^*u \neq 0 \), we have

\[
P_{\hat{\mu}, \bar{\Sigma}}(u^T X \leq u^T \bar{\mu}) = \frac{1}{2},
\]

\[
P_{\mu^*, \Sigma^*}(u^T X \leq u^T \bar{\mu}) = \Phi \left( \frac{u^T(\bar{\mu} - \mu^*)}{\sqrt{u^T \Sigma^* u}} \right).
\]

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

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

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

\[
0 \leq u^T(\bar{\mu} - \mu^*) \leq \sqrt{u^T \Sigma^* u}S_{1,a}d(\bar{\theta}, \theta^*).
\]  \( (78) \)

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 \( u^T(\bar{\mu} - \mu^*) \geq 0 \), we have

\[
\|\bar{\mu} - \mu^*\|_2 = \sup_{u \in \mathcal{U}_2} u^T (\bar{\mu} - \mu^*)
\]

\[
\leq S_{1,a}\|\Sigma^*\|_{op}^{1/2}d(\bar{\theta}, \theta^*).
\]
Similarly, let \( \mathcal{U}_\infty = \{ \pm e_j : j = 1, \ldots, p \} \), where \( e_j \) is a vector with \( j \)th coordinate being one and others being zero. By (78) with \( u \) restricted such that \( u \in \mathcal{U}_\infty \) and \( u^T (\bar{\mu} - \mu^*) \geq 0 \), we have

\[
\|\bar{\mu} - \mu^*\|_\infty = \sup_{u \in \mathcal{U}_\infty} u^T (\bar{\mu} - \mu^*) \\
\leq S_{1,a} \|\Sigma^*\|_{\text{max}}^{1/2} d(\bar{\theta}, \theta^*).
\]

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

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

\[
\text{TV}(P_{\bar{\mu}, \Sigma^*}, P_{\mu^*, \bar{\Sigma}}) \leq \text{TV}(P_{\bar{\mu}, \Sigma^*}, P_{\mu^*, \Sigma^*}) + \text{TV}(P_{\mu^*, \Sigma^*}, P_{\mu^*, \bar{\Sigma}}). \tag{79}
\]

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

\[
\text{TV}(P_{\bar{\mu}, \Sigma^*}, P_{\mu^*, \Sigma^*}) \leq \sqrt{2D_{\text{KL}}(P_{\bar{\mu}, \Sigma^*}||P_{\mu^*, \Sigma^*})} \\
\leq S_{1,a} d(\bar{\theta}, \theta^*). \tag{80}
\]

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

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

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

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

Combining (79) and (80) yields

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

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

\[
0 \leq P_{\mu, \Sigma} \left\{ (u^T X - u^T \bar{\mu})^2 \leq z_0 u^T \bar{\Sigma} u \right\} - P_{\mu, \bar{\Sigma}} \left\{ (u^T X - u^T \bar{\mu})^2 \leq z_0 u^T \bar{\Sigma} u \right\} \\
= P_{0, \Sigma} \left\{ (u^T X)^2 \leq z_0 u^T \bar{\Sigma} u \right\} - P_{0, \bar{\Sigma}} \left\{ (u^T X)^2 \leq z_0 u^T \bar{\Sigma} u \right\} \\
\leq d(\bar{\theta}, \theta^*) + S_{1,a} d(\bar{\theta}, \theta^*),
\]

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

\[
0 \leq P_{\mu, \Sigma} \left\{ (u^T X - u^T \bar{\mu})^2 \geq z_0 u^T \Sigma u \right\} - P_{\mu, \bar{\Sigma}} \left\{ (u^T X - u^T \bar{\mu})^2 \geq z_0 u^T \Sigma u \right\} \\
= P_{0, \Sigma} \left\{ (u^T X)^2 \geq z_0 u^T \Sigma u \right\} - P_{0, \bar{\Sigma}} \left\{ (u^T X)^2 \geq z_0 u^T \Sigma u \right\} \\
\leq d(\bar{\theta}, \theta^*) + S_{1,a} d(\bar{\theta}, \theta^*).
Notice that the choice of $z_0$ ensures that for $Z \in N(0,1)$,

$$P_{0,\Sigma} \left\{ (u^T X)^2 \leq z_0 u^T \Sigma u \right\} = P(Z^2 \leq z_0) = \frac{1}{2}$$

$$= P(Z^2 \geq z_0) = P_{0,\Sigma} \left\{ (u^T X)^2 \geq z_0 u^T \Sigma u \right\} .$$

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

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

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

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

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

$$\sqrt{\frac{u^T \Sigma u}{u^T \Sigma^* u}} - 1 \leq S_{2,a}(1 + S_{1,a})d(\bar{\theta}, \theta^*),$$

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

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

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

$$\|\Sigma - \Sigma^*\|_{\text{op}} = \sup_{u \in U_{2,\infty}} |u^T(\Sigma - \Sigma^*)u|$$

$$\leq 2S_{3,a}\|\Sigma^*\|_{\text{op}}d(\bar{\theta}, \theta^*) + S_{3,a}^2\|\Sigma^*\|_{\text{op}}(d(\bar{\theta}, \theta^*))^2,$$

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

To handle $\|\Sigma - \Sigma^*\|_{\text{max}}$, let $U_{2,\infty} = \{ \pm c_{ij} : i, j = 1, \ldots, p, i \neq j \}$, where $c_{ij}$ is a vector in $U_2$ with only $i$th and $j$th coordinates possibly being nonzero. For $u \in U_{2,\infty}$, we have $u^T \Sigma u = u_i^T \Sigma_i^T u_j$ and $u_i^T \Sigma^* u_j = u_i^T \Sigma^T u_j$, where $u_i \in \mathbb{R}^2$ is formed by $i$th and $j$th coordinates of $u$, and $\Sigma^*$ and $\Sigma$ are $2 \times 2$ matrices, formed by selecting $i$th and $j$th rows and columns from $\Sigma^*$ and $\Sigma$ respectively. Similarly as in the deviation of (83), applying inequality (82) with $u \in U_{2,\infty}$, we have

$$\|\Sigma_{ij} - \Sigma_{ij}^*\|_{\text{op}} = \sup_{u \in U_{2,\infty}} |u^T(\Sigma - \Sigma^*)u|$$

$$\leq 2S_{3,a}\|\Sigma_{ij}^*\|_{\text{op}}d(\bar{\theta}, \theta^*) + S_{3,a}^2\|\Sigma_{ij}^*\|_{\text{op}}(d(\bar{\theta}, \theta^*))^2.$$

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

$$\|\Sigma_{ij} - \Sigma_{ij}^*\|_{\text{max}}$$

$$\leq 4S_{3,a}\|\Sigma_{ij}^*\|_{\text{max}}d(\bar{\theta}, \theta^*) + 2S_{3,a}^2\|\Sigma_{ij}^*\|_{\text{max}}(d(\bar{\theta}, \theta^*))^2. \quad (84)$$
Taking the maximum on both sides of (84) over \( i \neq j \) gives the desired result:

\[
\|\bar{\Sigma} - \Sigma^*\|_{\text{max}} = \max_{i \neq j \in \{1,\ldots,p\}} \|\Sigma_{ij} - \Sigma^*_{ij}\|_{\text{max}} \\
\leq 4S_{3,a}\|\Sigma^*\|_{\text{max}}d(\bar{\theta}, \theta^*) + 2S_{3,a}^2\|\Sigma^*\|_{\text{max}}(d(\bar{\theta}, \theta^*))^2.
\]

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

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

\[
|P_{\mu^*, \Sigma^*}(u^TX = u^T\mu^*) - P_{\bar{\mu}, \bar{\Sigma}}(u^TX = u^T\mu^*)| \leq d(\bar{\theta}, \theta^*) < 1. \tag{85}
\]

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

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

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

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

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

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

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 \( u^T\Sigma^*u = 0 \). Hence the desired results hold by the remaining proofs in (i) and (ii). \( \blacksquare \)

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 \( \|\Sigma\|_{\text{max}} \leq M_1 \). For \( k \) fixed knots \( \xi_1, \ldots, \xi_k \) in \( \mathbb{R} \), denote \( \varphi(x) = (\varphi_1(x), \ldots, \varphi_k(x))^T \), 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 \( \varphi(X) - \mathbb{E}\varphi(X) \) 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 \( 1 - 2\delta \),

\[
\sup_{\|u\|_1 = 1} \left| u^T \left\{ \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) - \mathbb{E}\varphi(X) \right\} \right| \leq C_{\text{sp}11}M_1^{1/2} \sqrt{\frac{2\log(kp)}{n} + \log(\delta^{-1})},
\]
where \( C_{sp11} = \sqrt{2}C_{sg5} \), depending on the universal constant \( C_{sg5} \) in Lemma 69.

(iii) Let \( A_1 = \{ A \in \mathbb{R}^{kp \times kp} : \| A \|_{1,1} = 1 \} \). For any \( \delta > 0 \), we have that with probability at least \( 1 - 4\delta \), both the inequality in (ii) and

\[
\sup_{A \in A_1} \left\| \frac{1}{n} \sum_{i=1}^{n} \varphi^T(X_i)A\varphi_i - E\varphi^T(X)A\varphi(X) \right\| \leq C_{sp12}M_{11} \left\{ \sqrt{\frac{2\log(kp) + \log(\delta^{-1})}{n}} + \frac{2\log(kp) + \log(\delta^{-1})}{n} \right\},
\]

where \( M_{11} = M_{1}^{1/2}(M_{1}^{1/2} + \sqrt{2\pi}\|\xi\|_{\infty}) \), \( \|\xi\|_{\infty} = \max_{i=1, \ldots, k} |\xi_i| \), and \( C_{sp12} = \sqrt{2/\pi}C_{sp11} + C_{sx7}C_{sx6}C_{sx5} \). Constants \( (C_{sx5}, C_{sx6}, C_{sx7}) \) 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

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

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

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

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

\[
\frac{1}{n} \sum_{i=1}^{n} \varphi_i^T A \varphi_i - E\varphi^T A \varphi
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} (\varphi_i - E\varphi)^T A (\varphi_i - E\varphi) - E\{ (\varphi - E\varphi)^T A (\varphi - E\varphi) \} \tag{86}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} 2(E\varphi)^T A (\varphi_i - E\varphi). \tag{87}
\]

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

First, for \( A \in A_1 \), the term in (87) can be bounded as follows:

\[
\left\| 2(E\varphi)^T A \frac{1}{n} \sum_{i=1}^{n} \hat{\varphi}_i \right\| \leq 2 \| E\varphi \|_{\infty} \| A \|_{1,1} \left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\varphi}_i \right\|_{\infty} = 2 \| E\varphi \|_{\infty} \left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\varphi}_i \right\|_{\infty}
\]

\[
\leq 2 \left( M_{1}^{1/2} + \|\xi\|_{\infty} \right) \left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\varphi}_i \right\|_{\infty},
\]

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

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \right\|_\infty \leq 2 \log(kp) + \log(\delta^{-1}) \cdot \sqrt{\frac{\log(kp) + \log(\delta^{-1})}{n}}.
\]

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

\[
\sup_{A \in \mathcal{A}_1} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i^TA\tilde{\varphi}_i - \mathbb{E} \tilde{\varphi}_i^TA\tilde{\varphi}_i \right| \leq \sqrt{\frac{2}{\pi} C_{\text{sp}11} M_1^{1/2} \left( M_1^{1/2} + \sqrt{2\pi} \|\xi\|_\infty \right) \sqrt{\frac{2 \log(kp) + \log(\delta^{-1})}{n}}}.
\]

(88)

Next, notice that

\[
\sup_{A \in \mathcal{A}_1} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i^TA\tilde{\varphi}_i - \mathbb{E} \tilde{\varphi}_i^TA\tilde{\varphi}_i \right| = \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \otimes \tilde{\varphi}_i - \mathbb{E} \tilde{\varphi} \otimes \tilde{\varphi} \right\|_{\max}.
\]

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 \( C_{\text{ss}6}M_1 \). By Lemma 75, each element of the centered version, \( \tilde{\varphi}_i \otimes \tilde{\varphi}_i - \mathbb{E} \tilde{\varphi} \otimes \tilde{\varphi} \), is sub-exponential with tail parameter \( C_{\text{ss}6}C_{\text{ss}5}M_1 \). Then for any \( t > 0 \), by Lemma 76 and the union bound, we have that with probability at least \( 1 - 2k^2p^2e^{-t} \),

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \otimes \tilde{\varphi}_i - \mathbb{E} \tilde{\varphi} \otimes \tilde{\varphi} \right\|_{\max} \leq C_{\text{ss}7}C_{\text{ss}6}C_{\text{ss}5}M_1 \left( \sqrt{\frac{t}{n}} + \sqrt{\frac{t}{n}} \right).
\]

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

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \otimes \tilde{\varphi}_i - \mathbb{E} \tilde{\varphi} \otimes \tilde{\varphi} \right\|_{\max} \leq C_{\text{ss}7}C_{\text{ss}6}C_{\text{ss}5}M_1 \left( \sqrt{\frac{2 \log(kp) + \log(\delta^{-1})}{n}} + \sqrt{\frac{2 \log(kp) + \log(\delta^{-1})}{n}} \right).
\]

(89)

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

\[
\text{(90)}
\]

\[\text{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

\[
(1 - \epsilon) f'(t) - f\#(t) \leq -f'(1 - \epsilon) \epsilon.
\]

\text{(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

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

(iii) Suppose, in addition, that $f'(e^x)$ is concave in $u$ and $f^\#(e^x)$ 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 $h : \mathbb{R}^p \to \mathbb{R}$,

$$K_f(P_n, P_{\theta^*} ; h) \leq -f'(1 - \epsilon_1)\hat{\epsilon} + R_1|E_{P_{\theta^*},n}h(x) - E_{P_{\theta^*}}h(x)|,$$  

(91)

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) Notice that by definition,

$$(1 - \epsilon)f'(t) - f^\#(t) = (1 - \epsilon)f'(t) - tf'(t) + f(t)$$  

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

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

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

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

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

Combining the preceding three displays yields the desired result.

(ii) For any function $h$, we have

$$K_f(P_\epsilon, P_{\theta^*} ; h) = \epsilon E_Q f'(e^{h(x)}) + E_{P_{\theta^*}^*} \left\{(1 - \epsilon)f'(e^{h(x)}) - f^\#(e^{h(x)})\right\}$$

$$\leq E_{P_{\theta^*}} \left\{(1 - \epsilon)f'(e^{h(x)}) - f^\#(e^{h(x)})\right\},$$  

(92)

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

$$(1 - \epsilon)f'(e^{h(x)}) - f^\#(e^{h(x)}) \leq -f'(1 - \epsilon)\epsilon \leq -f'(1 - \epsilon_0)\epsilon.$$  

(93)

where $f'(1 - \epsilon) \geq f'(1 - \epsilon_0)$ for $\epsilon \leq \epsilon_0$ 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:

$$K_f(P_n, P_{\theta^*} ; h)$$

$$\leq \frac{1}{n} \sum_{i=1}^n U_i f'(e^{h(X_i)}) + \frac{1}{n} \sum_{i=1}^n (1 - U_i) f'(e^{h(X_i)}) - E_{P_{\theta^*}} f^\#(e^{h(x)})$$

$$\leq (1 - \hat{\epsilon})E_{P_{\theta^*},n} f'(e^{h(x)}) - E_{P_{\theta^*}} f^\#(e^{h(x)})$$  

(94)

$$\leq (1 - \hat{\epsilon})f'(e^{E_{P_{\theta^*},n}h(x)}) - f^\#(e^{E_{P_{\theta^*}}h(x)})$$

(95)

$$\leq -f'(1 - \epsilon_1)\hat{\epsilon} + |f^\#(e^{E_{P_{\theta^*},n}h(x)}) - f^\#(e^{E_{P_{\theta^*}}h(x)})|$$

(96)

$$\leq -f'(1 - \epsilon_1)\hat{\epsilon} + R_1|E_{P_{\theta^*},n}h(x) - E_{P_{\theta^*}}h(x)|.$$  

(97)
Line (94) follows because \( f'(t) \leq 0 \) for \( t > 0 \). Line (95) follows from Jensen’s inequality by the concavity of \( f'(e^u) \) and \( -f#(e^u) \) in \( u \). Line (96) follows because

\[
(1 - \hat{\epsilon}) f'(e^{E_{P_0^*},n}h(x)) - f#(e^{E_{P_0^*},n}h(x)) \leq -f'(1 - \epsilon_1) \hat{\epsilon},
\]

obtained by taking \( \epsilon = \hat{\epsilon} \) and \( t = e^{E_{P_0^*},n}h(x) \) in (90) and using \( f'(1 - \hat{\epsilon}) \geq f'(1 - \epsilon_1) \) for \( \hat{\epsilon} \leq \epsilon_1 \). Finally, line (97) follows because \( f#(e^u) \) is \( R_1 \)-Lipschitz in \( u \).

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

\[
K_f(P_n, P_{\theta^*}; h) \leq -f'(1 - \epsilon_1) \hat{\epsilon} + |E_{P_0^*},n f#(e^{h(x)}) - E_{P_0^*} f#(e^{h(x)})|.
\]

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

\[
(1 - \hat{\epsilon}) f'(e^{h(x)}) - f#(e^{h(x)}) \leq -f'(1 - \epsilon_1) \hat{\epsilon},
\]

which can be obtained by taking \( \epsilon = \hat{\epsilon} \) and \( t = e^{h(x)} \) in (90), similarly as (93). However, the bound (98) involves the moment difference of \( f#(e^{h(x)}) \) between \( P_{\theta} \) and \( P_{\theta^*},n \), which is difficult to control for \( h \) in our spline class, even with \( f#(e^u) \) Lipschitz in \( u \). In contrast, by exploiting the concavity of \( f'(e^u) \) and \( -f#(e^u) \) 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 \( \gamma \in \Gamma \),

\[
K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*}) \leq -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) + \text{pen}_1(\gamma)C_{\text{sp13}}R_1M_1\lambda_{11},
\]

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

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

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

\[
K_f(P_n, P_{\theta^*}; h_{\gamma,\mu^*})
\]

\[
\leq -f'(3/5)\hat{\epsilon} + R_1 \left| E_{P_{\theta^*},n} h_{\gamma,\mu^*}(x) - E_{P_{\theta^*}} h_{\gamma,\mu^*}(x) \right|
\]

\[
\leq -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) + R_1 \left| E_{P_{\theta^*},n} h_{\gamma}(x - \mu^*) - E_{P_{\theta^*}} h_{\gamma}(x) \right|.
\]

The last step (99) uses the fact that \( E_{P_{\theta^*}} h_{\gamma,\mu^*}(x) = E_{P_{\theta^*},\Sigma^*} h_{\gamma}(x) \) and \( E_{P_{\theta^*},n} h_{\gamma,\mu^*}(x) = E_{P_{\theta^*},n} h_{\gamma}(x - \mu^*) \), 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 \(\Omega_1\) holds, we have that \(\{X_i : U_i = 1, i = 1, \ldots, n\}\) are \(n_1\) independent and identically distributed observations from \(P_0^\ast\), where \(n_1 = \sum_{i=1}^{n}(1 - U_i) = n(1 - \epsilon) \geq (3/5)n\). Denote as \(\Omega_2\) the event that for any \(\gamma_1\) and \(\gamma_2\),

\[
\left| E_{P_0^\ast,n} \gamma_1^T \varphi(x - \mu^*) - E_{P(0,\Sigma^2)} \gamma_1^T \varphi(x) \right| \leq \|\gamma_1\|_1 C_{sp11} M_1^{1/2} \sqrt{\frac{2 \log(5p) + \log(\delta^{-1})}{(3/5)n}},
\]

and

\[
\left| E_{P_0^\ast,n} \gamma_2^T (\varphi(x - \mu^*) \otimes \varphi(x - \mu^*)) - E_{P(0,\Sigma^2)} \gamma_2^T (\varphi(x) \otimes \varphi(x)) \right|
\leq \|\gamma_2\|_1 C_{sp12} M_{11} \left\{ \sqrt{\frac{2 \log(5p) + \log(\delta^{-1})}{(3/5)n}} + \frac{2 \log(5p) + \log(\delta^{-1})}{(3/5)n} \right\},
\]

where \(C_{sp11}, C_{sp12},\) and \(M_{11}\) are defined as in Lemma 30 with \(\|\xi\|_\infty = 2\). In the event \(\Omega_2\), the preceding inequalities imply that for any \(\gamma = (\gamma_0, \gamma_1^T, \gamma_2^T) \in \Gamma\),

\[
\left| E_{P_0^\ast,n} h_\gamma(x - \mu^*) - E_{P(0,\Sigma^2)} h_\gamma(x) \right|
\leq \text{pen}_1(\gamma)(5/3)(C_{sp11} \vee C_{sp12}) M_{11} \lambda_{11},
\]

(100)

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

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 \(\gamma \in \Gamma\). \(\square\)

**Lemma 34** Suppose that \(X_1, \ldots, X_n\) are independent and identically distributed as \(X \sim P_0\). Let \(b > 0\) be fixed and \(g : \mathbb{R}^p \rightarrow [0, 1]^q\) be a vector of fixed functions. For a convex and twice differentiable function \(f : (0, \infty) \rightarrow \mathbb{R}\), define

\[
F(X_1, \ldots, X_n) = \sup_{\|w\|_1 = 1, \mu \in \mathbb{R}^p} \left\{ K_f(P_n, P_\theta; bw^T g_\mu) - K_f(P_\epsilon, P_\theta; bw^T g_\mu) \right\}
\]

\[
= \sup_{\|w\|_1 = 1, \mu \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} f'(e^{bw^T g_\mu}(X_i)) - Ef'(e^{bw^T g_\mu}(X)) \right\},
\]

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 \(g_{\mu,j} : \mathbb{R}^p \rightarrow [0, 1]\) denotes the \(j\)th component of \(g_\mu\). Then for any \(\delta > 0\), we have that with probability at least \(1 - 2\delta\),

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

where \(R_{2,b} = \sup_{|w| \leq b} \frac{d}{dn} f'(e^w)\) and \(C_{sg6}\) is the universal constant in Lemma 70.
Proof First, $F$ satisfies the bounded difference condition, because $|bw^Tg_\mu| \leq b$ with $\|w\|_1 = 1$ and $f'$ is non-decreasing by the convexity of $f$:

$$\sup_{X_1,\ldots,X_n} |F(X_1,\ldots,X_n) - F(X_1,\ldots,X'_n,\ldots,X_n)|$$

$$\leq \frac{f'(eb) - f'(e-b)}{n} \leq \frac{2bR_{2,b}}{n},$$

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

$$|F(X_1,\ldots,X_n) - EF(X_1,\ldots,X_n)| \leq 2bR_{2,b}t.$$

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

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

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

$$E \sup_{\|w\|_1=1,\mu \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n f'(e^{bw^Tg_\mu(X_i)}) - Ef'(e^{bw^Tg_\mu(X)}) \right\}$$

$$\leq 2E \sup_{\|w\|_1=1,\mu \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n \epsilon_i f'(e^{bw^Tg_\mu(X_i)}) \right\}$$

$$\leq 2R_{2,b}E \sup_{\|w\|_1=1,\mu \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n \epsilon_i bw^Tg_\mu(X_i) \right\}$$

$$\leq 2bR_{2,b}E \sup_{\mu \in \mathbb{R}^p} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_\mu(X_i) \right\|_\infty$$

$$\leq 2bR_{2,b}C_{\text{sg6}} \sqrt{V_g \log(2q)/n}. \quad (103)$$

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 $f'(e^u)$ is $R_{2,b}$-Lipschitz in $u \in [-b,b]$. Line (103) follows because

$$E \sup_{\mu \in \mathbb{R}^p} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_\mu(X_i) \right\|_\infty = E \sup_{\mu \in \mathbb{R}^p} \max_{j=1,\ldots,q} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_{\mu,j}(X_i) \right\|_\infty$$

$$= E \max_{j=1,\ldots,q} \sup_{\mu \in \mathbb{R}^p} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_{\mu,j}(X_i) \right\|_\infty$$

$$\leq C_{\text{sg6}} \sqrt{V_g \log(2q)/n}.$$

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 $E(\max_{j=1,\ldots,q} Z_{n,j}|X_1, \ldots, X_n) \leq C_{\text{sg6}} \sqrt{q} \log(2q)/n$, and then $E(\max_{j=1,\ldots,q} Z_{n,j}) \leq C_{\text{sg6}} \sqrt{q} \log(2q)/n$.

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

\textbf{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 $K_{\text{HG}}$. This is true because $f'(e^u)$ will be replaced by identity function which is just 1-Lipschitz.

\textbf{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

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

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

\textbf{Proof} First, $K_f(P^*, P_{\hat{\theta}}; h)$ can be bounded as

$$K_f(P^*, P_{\hat{\theta}}; h) = \epsilon E_Q f'(e^{h(x)}) + (1 - \epsilon) E_{P^*} f'(e^{h(x)}) - E_{P_{\hat{\theta}}} f^#(e^{h(x)})$$

$$\geq f'(e^{-b})\epsilon + E_{P^*} f'(e^{h(x)}) - E_{P_{\hat{\theta}}} f^#(e^{h(x)})$$

$$= f'(e^{-b})\epsilon + K_f(P^*, P_{\hat{\theta}}; h),$$

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

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

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

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

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 $R_{3,b}$.

\textbf{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 $\gamma \in \Gamma_p$ with $\gamma_0 = 0$ and $\text{pen}_1(\gamma) = b_1$,

$$K_f(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\theta}}) \geq f''(1) \left\{ E_{P_n} h_{\gamma, \hat{\theta}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\theta}}(x) \right\} + f'(e^{-b_1})\epsilon - \frac{1}{2} b_1^2 R_{3,b_1} - b_1 R_{2,b_1} \lambda_{12}$$
where, with $C_{\text{rad4}} = C_{\text{sg6}}C_{\text{rad3}},$

$$
\lambda_{12} = C_{\text{rad4}}\sqrt{\frac{4\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}},
$$
depending on the universal constants $C_{\text{sg6}}$ and $C_{\text{rad3}}$ in Lemma 70 and Corollary 82.

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

$$
h_{r_p,\beta,c}(x) = \beta_0 + \sum_{j=1}^p \beta_{1j} \text{ramp}(x_j - c_j) + \sum_{1 \leq i \neq j \leq p} \beta_{2,ij} \text{ramp}(x_i)\text{ramp}(x_j),
$$

where $c = (c_1, \ldots, c_p)^T$ with $c_j \in \{0, 1\},$ and $\beta = (\beta_0, \beta_1^T, \beta_2^T)^T$ with $\beta_1 = (\beta_{1j} : j = 1, \ldots, p)^T$ and $\beta_2 = (\beta_{2,ij} : 1 \leq i \neq j \leq p).$ Then for any $\gamma \in \Gamma_{rp}$ with $\gamma_0 = 0$ and $\text{pen}_1(\gamma) = b_1,$ we have $\beta_0 = 0$ and $\text{pen}_1(\beta) = b_1$ correspondingly, and hence $h_\gamma(x) = h_{r_p,\beta,c}(x) \in [-b_1, b_1]$ by the boundedness of the ramp function in $[0, 1].$ Moreover, $h_{r_p,\beta,c}(x)$ with $\beta_0 = 0$ and $\text{pen}_1(\beta) = b_1$ can be expressed in the form $b_1w^T g(x),$ 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 $\text{ramp}(x_j)$ and $\text{ramp}(x_j - 1)$ for $j = 1, \ldots, p,$ and $\text{ramp}(x_i)\text{ramp}(x_j)$ for $1 \leq i \neq j \leq p.$ For symmetry, $\text{ramp}(x_i)\text{ramp}(x_j)$ and $\text{ramp}(x_j)\text{ramp}(x_i)$ 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 $\text{ramp}(x_j)$ and $\text{ramp}(x_j - 1)$ is nonzero for each $j,$ but this property is not used in the subsequent discussion.

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

$$
K_f(P_n, P_{\hat{\rho}}; h_{\gamma,\hat{\mu}}) 
\geq K_f(P_n, P_{\hat{\rho}}; h_{\gamma,\hat{\mu}}) - \{K_f(P_n, P_{\hat{\rho}}; h_{\gamma,\hat{\mu}}) - K_f(P_n, P_{\hat{\rho}}; h_{\gamma,\hat{\mu}})\}.
$$

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

$$
K_f(P_n, P_{\hat{\rho}}; h_{\gamma,\hat{\mu}}) 
\geq f''(1) \left\{ E_{\rho_\gamma} h_{\gamma,\hat{\mu}}(x) - E_{\rho_\gamma} h_{\gamma,\hat{\mu}}(x) \right\} + f'(e^{-b_1})e - \frac{1}{2} b_1^2 R_{3,b_1}.
$$

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 $\gamma \in \Gamma_{rp}$ with $\gamma_0 = 0$ and $\text{pen}_1(\gamma) = b_1,$

$$
\{K_f(P_n, P_{\hat{\rho}}; h_{\gamma,\hat{\mu}}) - K_f(P_n, P_{\hat{\rho}}; h_{\gamma,\hat{\mu}})\} 
\leq b_1 R_{2,b_1} \left\{ C_{\text{sg6}} \sqrt{\frac{V_{\rho_\gamma} \log(2q)}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}} \right\}
\leq b_1 R_{2,b_1} \left\{ C_{\text{sg6}} C_{\text{rad3}} \sqrt{\frac{4\log(2p(p+1))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}} \right\},
$$

where $V_{\rho_\gamma} = 4C_{\text{rad3}}^2$ 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 $\mu = (\mu_1, \ldots, \mu_p)^T$ and, as defined in Lemma 34,
The sub-gaussian random variable $Z_j$ is a product of moving-knot ramp functions, $\text{ramp}(x)$ for $1 \leq j_1 \neq j_2 \leq p$. By Lemma 60, the VC index of moving-knot ramp functions is 2. By applying Corollary 82 (i) and (ii) with vanishing $H$, 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 $C_{\text{rad}3} \sqrt{4/n}$ 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 $0 \leq \sigma_1, \sigma_2 \leq M^{1/2}$, denote $D_h = \text{E}h(\sigma_1 Z + \delta) - \text{E}h(\sigma_2 Z)$, where $h$ is a function on $\mathbb{R}$ and $Z$ is a standard Gaussian random variable. For $h_1(x) = \pm \text{ramp}(x)$, if $|D_{h_1}| \leq a$ for $a \in (0, 1/2)$, then we have

$$|\delta| \leq S_{4,a}|D_{h_1}|,$$

(104)

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

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

(105)

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

**Remark 39** Define a ramp function class

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

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

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

Then we have

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

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, $D \leq 1/3$.

**Proof**

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

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

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

---

67
(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 $g'(t) = \frac{1}{2}E\{1_{\sigma_1 Z + t \in [-1, 1]}\} \geq 0$. Moreover, $g'(t)$ is non-increasing: for $0 \leq t_1 < t_2$,

\[
g'(t_1) = \frac{1}{2}P(-1 - t_1 \leq \sigma_1 Z \leq 1 - t_1)
\]

\[= \frac{1}{2}P(-1 - t_1 \leq \sigma_1 Z \leq 1 - t_2) + \frac{1}{2}P(1 - t_2 \leq \sigma_1 Z \leq 1 - t_1),
\]

\[
g'(t_2) = \frac{1}{2}P(-1 - t_2 \leq \sigma_1 Z \leq 1 - t_2)
\]

\[= \frac{1}{2}P(-1 - t_1 \leq \sigma_1 Z \leq 1 - t_2) + \frac{1}{2}P(-1 - t_2 \leq \sigma_1 Z \leq -1 - t_1),
\]

and

\[
P(1 - t_2 \leq \sigma_1 Z \leq 1 - t_1) = P(t_1 - 1 \leq \sigma_1 Z \leq t_2 - 1)
\]

\[\leq P(t_1 + 1 \leq \sigma_1 Z \leq t_2 + 1) = P(-1 - t_2 \leq \sigma_1 Z \leq -1 - t_1).
\]

The last inequality holds because $(t - 1)^2 \leq (t + 1)^2$ for any $t \geq 0$ and hence $\int_{t_1}^{t_2} \exp\{-(t - 1)^2/(2\sigma_1^2)\} dt \geq \int_{t_1}^{t_2} \exp\{-(t + 1)^2/(2\sigma_1^2)\} dt$. To show (ii), we write $g(t) = g(t) - g(0) = t \int_0^1 g'(t) \, dz$. Then $g(t)/t = \int_0^1 g'(t) \, dz$ is non-increasing in $t$ because $g'$ is non-increasing.

To show (iii), we notice that $h(x) \geq 1_{x > 1}$ and hence for $t \geq 1$,

\[
g(t) + \frac{1}{2} = Eh(\sigma_1 Z + t)
\]

\[\geq 1 - P(\sigma_1 Z + t \leq 1) = 1 - P(\sigma_1 Z \geq t - 1)
\]

\[\geq 1 - \exp\{-(t - 1)^2/(2\sigma_1^2)\}.
\]

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

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

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 $\sigma_3 \leq M^{1/2}$ and the definition of $S_{4,a}$.

When $\delta$ is negative, a similar argument taking $h(x) = \text{ramp}(x)$, which is the same as $\text{ramp}(-x) - 1$, 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 \( h(x) = \text{ramp}(x - 1) \). Notice that \( h(x) \) is \( (1/2) \)-Lipschitz and hence \( |h(x + \delta) - h(x)| \leq |\delta|/2 \). Then by the triangle inequality, we have

\[
Eh(\sigma_1 Z) - Eh(\sigma_2 Z) \leq D_h + |\delta|/2.
\]

Define \( g(t) = Eh(tZ) - Eh(\sigma_2 Z) \) for \( t \geq 0 \). Then \( g(\sigma_2) = 0 \) and \( g(\sigma_1) \leq D_h + |\delta|/2 \). The derivative \( g'(t) = \frac{1}{2} E\{Z_{1 + tZ \in [0,2]} \} \) can be calculated as

\[
g'(t) = \frac{1}{2} \int_0^{2/t} e^{-z^2/2} dz = \frac{1}{2} \int_0^{2/t} \left( 1 - e^{-z^2/2} \right) dz.
\]

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

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

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

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

\[\Box\]

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

\[
|\tilde{\rho} \delta_1 \tilde{\sigma}_2 - \rho \sigma_1 \sigma_2| \leq M^{1/2} \|\tilde{\sigma} - \sigma\|_1 + S_7(\|D_h\| + \Delta/2),
\]

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

**Proof** First, we handle the effect of different means and standard deviations between \( \tilde{X} \) and \( X \) in \( D_h \). Denote \( D_h^1 = Eh(\tilde{Y}) - Eh(X) \), where \( \tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2)^T \) is a Gaussian random vector with mean 0 and variance matrix \( \begin{pmatrix} \sigma_2^2 & \sigma_1 \sigma_2 \rho \\ \sigma_1 \sigma_2 \rho & \sigma_2^2 \end{pmatrix} \). Then we have

\[
|D_h^1| \leq |D_h| + \Delta/2.
\]

In fact, assume that \( \tilde{Y} = (\sigma_1 Z_1, \sigma_2 Z_2)^T \) and \( X = \delta + (\tilde{\sigma}_1 Z_1, \tilde{\sigma}_2 Z_2)^T \), where \( (Z_1, Z_2)^T \) is a Gaussian random vector with mean 0 and variance matrix \( \begin{pmatrix} 1 & \tilde{\rho} \\ \tilde{\rho} & 1 \end{pmatrix} \). For \( h(x) =
Substituting this into the expression for $Eh(\tilde{Y}) - Eh(X)$, we have
\[
|Eh(\tilde{Y}) - Eh(X)| \\
\leq |Eh(\tilde{X}) - Eh(X)| + \sum_{j=1,2} E |\text{ramp}(\tilde{Y}_j) - \text{ramp}(\tilde{X}_j)| \\
\leq |Eh(\tilde{X}) - Eh(X)| + \sum_{j=1,2} \{|\delta_j^2 + (\tilde{\sigma}_j - \sigma_j)^2|/2 \}^{1/2} \\
\leq |Eh(\tilde{X}) - Eh(X)| + \Delta/2.
\]

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

Next, we show that
\[
|\tilde{\rho} - \rho| \sigma_1 \sigma_2 \leq 8\pi Me^{1/(4M)}|D^4_h|.
\]

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 $g(t) = Eh(Y) - Eh(X)$ for $t \in [-1, 1]$, where $Y = (Y_1, Y_2)^T$ 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 $g(0) = 0$ and $g(\tilde{\rho}) = D^4_h$. The derivative $g'(t)$ can be calculated as
\[
g'(t) = \frac{1}{4} \sigma_1 \sigma_2 E[I_{Y_1 \in [-1, 1]}] I_{Y_2 \in [-1, 1]}.
\]

In fact, $Y_1$ can be represented as $Y_1 = t_{\sigma_1} Y_2 + \sqrt{1 - t^2} \sigma_1 Z_1$, where $Z_1$ is a standard Gaussian variable independent of $Y_2$. Then direct calculation yields
\[
g'(t) = \frac{1}{2} E[I_{Y_1 \in [-1, 1]}] \left( \frac{\sigma_1}{\sigma_2} Y_2 - \frac{t}{\sqrt{1 - t^2}} \sigma_1 Z_1 \right) \text{ramp}(Y_2) \\
= \frac{1}{2} E[I_{Y_1 \in [-1, 1]}] \left( \frac{\sigma_1}{\sigma_2} Y_2 - \frac{t}{1 - t^2} Y_1 \right) \text{ramp}(Y_2) \\
= \frac{1}{2} E[I_{Y_1 \in [-1, 1]}] \left( \frac{1}{1 - t^2} \frac{\sigma_1}{\sigma_2} Y_2 - \frac{t}{1 - t^2} Y_1 \right) \text{ramp}(Y_2) \\
= \frac{1}{2} E[I_{Y_1 \in [-1, 1]}] \left( Y_2 - t \frac{\sigma_2}{\sigma_1} Y_1 \right) \text{ramp}(Y_2).
\]

By Stein’s lemma using the fact that $Y_2$ given $Y_1$ is Gaussian with mean $t_{\sigma_1} Y_1$ and variance $(1 - t^2)\sigma_2^2$, we have
\[
E \left\{ \left( Y_2 - t_{\sigma_1} Y_1 \right) \text{ramp}(Y_2) \Big| Y_1 \right\} = E \left\{ (1 - t^2)\sigma_2^2 Y_2 \Big| Y_2 \in [-1, 1] \right\} Y_1.
\]

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 $\sigma_1 \leq \sigma_2$, because formula (108) is symmetric in $Y_1$ and $Y_2$. By the previous representation of $Y_1$, we have $|Y_1| \leq \sqrt{1 - t^2}\sigma_1 |Z_1| + \frac{\sigma_1}{\sigma_2} |Y_2| \leq \sigma_1 |Z_1| + |Y_2|$ and hence

$$g'(t) \geq \frac{1}{4} \sigma_1 \sigma_2 \mathbb{E} [Y_1 \in [-1,1] Y_2 \in [-1/2,1/2]]$$

$$\geq \frac{1}{4} \sigma_1 \sigma_2 \mathbb{E} [\sigma_1 |Z_1| \in [-1/2,1/2] Y_2 \in [-1/2,1/2]]$$

$$\geq S_7^{-1} \sigma_1 \sigma_2,$$

where $S_7 = 4 \{ (\frac{1}{\sqrt{2\pi M}} e^{-1/(8M)}) \lor (1 - 2e^{-1/(8M)}) \}^{-2}$. The last inequality follows by the independence of $Z_1$ and $Y_2$, $\sigma_1 \leq M^{1/2}$, $\sigma_2 \leq M^{1/2}$, and the two probability bounds:

$$P(\mathcal{M}^{1/2}|Z_1| \leq 1/2) \geq \frac{1}{\sqrt{2\pi M}} e^{-(1/2)^2/(2M)}$$

and

$$P(\mathcal{M}^{1/2}|Z_1| \leq 1/2) \geq 1 - 2e^{-(1/2)^2/(2M)}.$$  

Hence by the mean value theorem, we have

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

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

$$(\hat{\rho} - \rho) \sigma_1 \sigma_2 \leq S_7 g(\hat{\rho}) = S_7 D_h^1.$$

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

Finally, by the triangle inequality, we find

$$|\hat{\rho}\hat{\sigma}_1 \hat{\sigma}_2 - \rho \sigma_1 \sigma_2| \leq |\hat{\sigma}_1 \hat{\sigma}_2 - \sigma_1 \sigma_2| + |\hat{\rho} - \rho| \sigma_1 \sigma_2$$

$$\leq M^{1/2}(|\hat{\sigma}_1 - \sigma_1| + |\hat{\sigma}_2 - \sigma_2|) + |\hat{\rho} - \rho| \sigma_1 \sigma_2.$$  

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

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

$$\mathcal{R}_2 = \{ \pm \text{ramp}(x_j - c), (x_1, x_2) \in \mathbb{R}^2 : j = 1, 2, c = 0, 1 \} \cup \{ \pm \text{ramp}(x_1)\text{ramp}(x_2), (x_1, x_2) \in \mathbb{R}^2 \}.$$

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

$$D \overset{\text{def}}{=} \sup_{h \in \mathcal{R}_2} \left\{ \mathbb{E} h(\hat{X}) - \mathbb{E} h(X) \right\} \leq a.$$  

Then combining Lemma 38–40 yields

$$\max(|\delta_1|, |\delta_2|) \leq S_{4,a} D,$$

$$\max(|\hat{\sigma}_1 - \sigma_1|, |\hat{\sigma}_2 - \sigma_2|) \leq S_{6,a} D,$$

$$\max(|\hat{\sigma}_1^2 - \sigma_1^2|, |\hat{\sigma}_2^2 - \sigma_2^2|, |\hat{\rho}\hat{\sigma}_1 \hat{\sigma}_2 - \rho \sigma_1 \sigma_2|) \leq S_{8,a} D$$

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 $a \in (0, 1/2)$,

$$D \overset{\text{def}}{=} \sup_{\gamma \in \Gamma_{\text{rp}}, \text{pen}_1(\gamma) = 1} \left\{ E_{P_0} h_{\gamma, \hat{\mu}}(x) - E_{P_0} h_{\gamma, \hat{\mu}}(x) \right\} \leq a. \quad (109)$$

Then we have

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

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

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

Proof For any $\gamma \in \Gamma_{\text{rp}}$ with $\text{pen}_1(\gamma) = 1$, the function $h_{\gamma, \hat{\mu}}(x) = h_{\gamma}(x - \hat{\mu})$ can be expressed as $h_{\gamma, \hat{\mu}, c}(x - \hat{\mu})$ with $\text{pen}_1(\beta) = 1$. 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

$$\max \left( |\hat{\Sigma}_{ii} - \Sigma_{ii}^*|, |\hat{\Sigma}_{jj} - \Sigma_{jj}^*|, |\hat{\Sigma}_{ij} - \Sigma_{ij}^*| \right) \leq S_{8,a} D,$$
(iii) Let $A_2 = \{ A \in \mathbb{R}^{kp \times kp} : \|A\|_F = 1, A^T = A \}$. For any $\delta > 0$, we have that with probability at least $1 - 3\delta$, both the inequality in (ii) and

$$
\frac{1}{\sqrt{kp}} \sup_{A \in A_2} \left| \frac{1}{n} \sum_{i=1}^{n} \varphi^T(X_i) A \varphi(X_i) - E \varphi^T(X) A \varphi(X) \right| 
\leq C_{sp2k} kM_{21} \left\{ \sqrt{\frac{kp + \log(\delta^{-1})}{n}} + \sqrt{\frac{kp}{n}} \right\},
$$

where $M_{21} = M_2^{1/2}(M_2^{1/2} + \sqrt{2\pi}\|\xi\|_{\infty})$, $\|\xi\|_{\infty} = \max_{l=1,...,k} |\xi_l|$, $C_{sp2k} = \sqrt{2/\pi}C_{sp21} + C_{sg8}$, and $C_{sg8}$ is the universal constant in Lemma 72.

**Proof** (i) First, we show that $w^T \varphi(x)$ 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 $|w^T(\varphi(x_1) - \varphi(x_2))| \leq \sum_{l=1}^{k} w_l \|\varphi_l(x_1) - \varphi_l(x_2)\|_2 \leq \sum_{l=1}^{k} w_l \|x_1 - x_2\|_2 \leq k^{1/2} \|x_1 - x_2\|_2$, where $w$ is partitioned as $w = (w_1^T, \ldots, w_k^T)^T$, 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 $\varSigma_{1/2} Z$, 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

$$
|w^T \varphi(\Sigma_{1/2} z_1) - w^T \varphi(\Sigma_{1/2} z_2)| 
\leq k^{1/2} \|\Sigma_{1/2}(z_1 - z_2)\|_2 \leq k^{1/2} \|\Sigma_{1/2}\|_{op} \|z_1 - z_2\|_2 \leq (kM_2)^{1/2} \|z_1 - z_2\|_2.
$$

Hence $w^T \varphi(X)$ 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 $t > 0$,

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

That is, $w^T(\varphi(X) - E \varphi(X))$ 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, $w^T(\varphi(X) - E \varphi(X))$ is sub-gaussian with tail parameter $(kM_2)^{1/2}$ for any $L_2$ unit vector $w$. Then $w^T \{ n^{-1} \sum_{i=1}^{n} \varphi(X_i) - E \varphi(X) \}$ is sub-gaussian with tail parameter $C_{sg5}(kM_2/n)^{1/2}$ by sub-gaussian concentration (Lemma 69). Hence by definition, we have that $n^{-1} \sum_{i=1}^{n} \varphi(X_i) - E \varphi(X)$ is a sub-gaussian random vector with tail parameter $C_{sg5}(kM_2/n)^{1/2}$. Notice that

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

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

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) - E \varphi(X) \right\|_2 \leq C_{sg7} C_{sg5}(kM_2)^{1/2} \left\{ \sqrt{\frac{kp}{n}} + \sqrt{\frac{\log(\delta^{-1})}{n}} \right\} 
\leq \sqrt{2} C_{sg7} C_{sg5}(kM_2)^{1/2} \sqrt{\frac{kp + \log(\delta^{-1})}{n}}.
$$
We handle the concentration of the two terms separately. Denote $\varphi_i = \varphi(X_i)$, $\varphi = \varphi(X)$, $\tilde{\varphi}_i = \varphi_i - E\varphi$, and $\tilde{\varphi} = \varphi - E\varphi$.

First, for $A \in A_2$ the term in (111) can be bounded as follows:

$$
\left| 2(E\varphi)^T A \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \right| \leq 2\|E\varphi\|_2 \|A\|_{op} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \right| \leq 2\|E\varphi\|_2 \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \right|_2 
$$

$$
\leq 2\sqrt{kp} \left( \frac{M_2^{1/2}}{\sqrt{2\pi}} + \|\xi\|_\infty \right) \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \right|_2 ,
$$

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

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

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

$$
\frac{1}{\sqrt{kp}} \sup_{A \in A_2} \left| 2(E\varphi)^T A \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \right| \leq \frac{\sqrt{2}}{\pi} C_{sp21}(kM_2)^{1/2} \left( \frac{M_2^{1/2}}{\sqrt{2\pi}} + \|\xi\|_\infty \right) \sqrt{\frac{kp + \log(\delta^{-1})}{n}} .
$$

Next, consider an eigen-decomposition $A = \sum_{l=1}^{kp} \lambda_l w_l w_l^T$, where $\lambda_l$’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:

$$
\sup_{A \in A_2} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i^T A \tilde{\varphi}_i - E\tilde{\varphi}_i^T A\tilde{\varphi} \right| = \sup_{A \in A_2} \left| \sum_{l=1}^{kp} \lambda_l w_l^T \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \tilde{\varphi}_i^T - E\tilde{\varphi} \tilde{\varphi}^T \right) w_l \right| 
$$

$$
\leq \sup_{A \in A_2} \left( \sum_{l=1}^{kp} |\lambda_l| \right) \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \tilde{\varphi}_i^T - E\tilde{\varphi} \tilde{\varphi}^T \right|_{op} \leq \sqrt{kp} \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \tilde{\varphi}_i^T - E\tilde{\varphi} \tilde{\varphi}^T \right\|_{op} .
$$

74
The last inequality uses the fact that \( \|A\|_F = (\sum_{i=1}^{kp} \lambda_i^2)^{1/2} = 1 \) and hence \( \sum_{i=1}^{kp} |\lambda_i| \leq \sqrt{kp} \) for \( A \in A_2 \). 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 \( 1 - 2\delta \),

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i \tilde{\varphi}_i^T - \mathbb{E} \tilde{\varphi} \tilde{\varphi}^T \right\| \leq C_{sgs} k M_2 \left\{ \sqrt{\frac{kp + \log(\delta^{-1})}{n}} + \frac{kp + \log(\delta^{-1})}{n} \right\},
\]

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

\[
\frac{1}{\sqrt{kp}} \sup_{A \in A_2} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}_i A \tilde{\varphi}_i - \mathbb{E} \tilde{\varphi} A \tilde{\varphi} \right| \leq C_{sgs} k M_2 \left\{ \sqrt{\frac{kp + \log(\delta^{-1})}{n}} + \frac{kp + \log(\delta^{-1})}{n} \right\}.
\]

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 \( \gamma = (\gamma_0, \gamma_1, \gamma_2)^T \in \Gamma \),

\[
K_f(P_n, P_{\theta^*}; h_{\gamma, \mu^*}) \leq -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) + \text{pen}_2(\gamma_1)(5/3)C_{sp_{21}} M_2^{1/2} R_1 \lambda_{21} + \text{pen}_2(\gamma_2)(25\sqrt{5}/3)C_{sp_{22}} M_2 R_1 \sqrt{\nu} \lambda_{31},
\]

where \( C_{sp_{21}} \) and \( C_{sp_{22}} \) are defined as in Lemma 43, \( M_2 = M_2^{1/2} (M_2^{1/2} + 2\sqrt{2\pi}) \), 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 \( \Omega_1 = \{ |\hat{\epsilon} - \epsilon| \leq \sqrt{\epsilon(1 - \epsilon)/(n\delta)} \} \). By Chebyshev’s inequality, we have \( P(\Omega_1) \geq 1 - \delta \). In the event \( \Omega_1 \), we have \( |\hat{\epsilon} - \epsilon| \leq 1/5 \) by the assumption \( \sqrt{\epsilon(1 - \epsilon)/(n\delta)} \leq 1/5 \) and hence \( \hat{\epsilon} \leq 2/5 \) by the assumption \( \epsilon \leq 1/5 \). By Lemma 31 with \( \epsilon_1 = 2/5 \), it holds in the event \( \Omega_1 \) that for any \( \gamma \in \Gamma \),

\[
K_f(P_n, P_{\theta^*}; h_{\gamma, \mu^*})
\]

\[
\leq -f'(3/5)\hat{\epsilon} + R_1 \left| E_{P_{\theta^*}} h_{\gamma, \mu^*}(x) - E_{P_{\theta^*}} h_{\gamma, \mu^*}(x) \right|
\]

\[
\leq -f'(3/5)(\epsilon + \sqrt{\epsilon/(n\delta)}) + R_1 \left| E_{P_{\theta^*}} h_{\gamma}(x - \mu^*) - E_{P_{0,\Sigma^*}} h_{\gamma}(x) \right|.
\]

The last step also uses the fact that \( E_{P_{\theta^*}} h_{\gamma, \mu^*}(x) = E_{P_{0,\Sigma^*}} h_{\gamma}(x) \) and also \( E_{P_{\theta^*}} h_{\gamma, \mu^*}(x) = E_{P_{\theta^*}} h_{\gamma}(x - \mu^*) \), 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 \( \Omega_1 \) holds, we have that \( \{ X_i : U_i = 1, i = 1, \ldots, n \} \) are \( n_1 \) independent and identically
distributed observations from $P_0$, where $n_1 = \sum_{i=1}^n (1 - U_i) = n(1 - \epsilon) \geq (3/5)n$. Denote as $\Omega_2$ the event that for any $\gamma_1$ and $\gamma_2$,

$$
\left| E_{P_0, \gamma_1} \varphi(x - \mu^*) - E_{P_0, \gamma_2} \varphi(x) \right| \leq \| \gamma_1 \|_2 C_{sp21} M_1^{1/2} \sqrt{\frac{5p + \log(\delta^{-1})}{(3/5)n}},
$$

and

$$
\left| E_{P_0, \gamma_1} \varphi(x - \mu^*) \otimes \varphi(x - \mu^*) - E_{P_0, \gamma_2} \varphi(x) \otimes \varphi(x) \right|
\leq \| \gamma_2 \|_2 C_{sp22} M_2^{1/2} \sqrt{\frac{5p + \log(\delta^{-1})}{(3/5)n} + \frac{5p + \log(\delta^{-1})}{(3/5)n}},
$$

where $C_{sp21}$, $C_{sp22}$, and $M_2$ are defined as in Lemma 43 with $\| \xi \|_\infty = 1$. In the event $\Omega_2$, the preceding inequalities imply that for any $\gamma = (\gamma_0, \gamma_1, \gamma_2)^T \in \Gamma$,

$$
\left| E_{P_0, \gamma} h_{\gamma}(x - \mu^*) - E_{P_0, \gamma} h_{\gamma}(x) \right|
\leq \text{pen}_2(\gamma_1)(5/3) C_{sp21} M_1^{1/2} \lambda_1 + \text{pen}_2(\gamma_2)(5/3) C_{sp22} M_2^{1/2} \lambda_2 + \lambda_1 5, \quad (115)
$$

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

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 $\gamma \in \Gamma$. \hfill \blacksquare

**Lemma 45** Suppose that $X_1, \ldots, X_n$ are independent and identically distributed as $X \sim P_0$. Let $b > 0$ be fixed and $g : \mathbb{R}^p \rightarrow [0, 1]^q$ be a vector of fixed functions. For a convex and twice differentiable function $f : (0, \infty) \rightarrow \mathbb{R}$, define

$$
F(X_1, \ldots, X_n) = \sup_{\| \xi \|_2 = 1, \mu \in \mathbb{R}^p, \eta_0 \in [0, 1]^q} \left\{ K_f(P_{n, \mu}; bw^T g_{\mu, \eta_0}) - K_f(P_{0, \mu}; bw^T g_{0, \eta_0}) \right\}
$$

$$
= \sup_{\| \xi \|_2 = 1, \mu \in \mathbb{R}^p, \eta_0 \in [0, 1]^q} \left\{ \frac{1}{n} \sum_{i=1}^n f'(e^{bw^T g_{\mu, \eta_0}}(X_i)) - E f'(e^{bw^T g_{0, \eta_0}}(X)) \right\},
$$

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 $\sqrt{V_n} / n$ for $j = 1, \ldots, q$, where $(\epsilon_1, \ldots, \epsilon_n)$ are Rademacher variables, independent of $(X_1, \ldots, X_n)$, and $g_{\mu, \eta_0,j} : \mathbb{R}^p \rightarrow [-1, 1]$ denotes the $j$th component of $g_{\mu, \eta_0}$. Then for any $\delta > 0$, we have that with probability at least $1 - 2\delta$,

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

where $R_{2,b^2} = \sup_{|u| \leq b^2} \frac{d}{du} f'(u)$ and $C_{sg,12}$ is the universal constant in Lemma 67.
Proof First, \( F \) satisfies the bounded difference condition, because \( |bw^T g_{\mu,\eta_0}| \leq b\sqrt{q} \) with \( \|w\|_2 \leq 1 \) and \( f' \) is non-decreasing by the convexity of \( f \):

\[
\sup_{X_1, \ldots, X_n, X'_i} |F(X_1, \ldots, X_n) - F(X_1, \ldots, X'_i, \ldots, X_n)| \leq \frac{f'(e^{b\sqrt{q}}) - f'(e^{-b\sqrt{q}})}{n} \leq \frac{2b\sqrt{q}R_{2,b}}{n},
\]

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

\[
|F(X_1, \ldots, X_n) - EF(X_1, \ldots, X_n)| \leq 2b\sqrt{q}R_{2,b}t.
\]

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

\[
|F(X_1, \ldots, X_n) - EF(X_1, \ldots, X_n)| \leq bR_{2,b} \sqrt{\frac{2q\log(\delta^{-1})}{n}}.
\]

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

\[
E \sup_{\|w\|_2=1, \mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} \left\{ \frac{1}{n} \sum_{i=1}^n f'(e^{bw^T g_{\mu,\eta_0}(X_i)}) - EF'(e^{bw^T g_{\mu,\eta_0}(X)}) \right\}
\]

\[
\leq 2E \sup_{\|w\|_2=1, \mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} \left\{ \frac{1}{n} \sum_{i=1}^n \epsilon_i f'(e^{bw^T g_{\mu,\eta_0}(X_i)}) \right\}
\]

\[
\leq 2R_{2,b\sqrt{q}} \sqrt{\frac{2qV_g}{n}}.
\]

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 \( f'(e^t) \) is \( R_{2,b\sqrt{q}} \)-Lipschitz in \( u \in [-b\sqrt{q}, b\sqrt{q}] \). Line (118) follows because

\[
E \sup_{\mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_{\mu,\eta_0}(X_i) \right\|_2 \leq \left\{ E \sup_{\mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_{\mu,\eta_0}(X_i) \right\|_2^2 \right\}^{1/2}
\]

\[
\leq \left\{ \sum_{j=1}^q E \sup_{\mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_{\mu,\eta_0,j}(X_i) \right\|_2^2 \right\}^{1/2}
\]

\[
\leq C_{sg,12} \sqrt{\frac{2qV_g}{n}}.
\]
For the last step, we use the assumption that conditionally on \((X_1, \ldots, X_n)\), the random variable \(Z_{p,i} = \sup_{\mu \in \mathbb{E}, \rho \in [0,1]} |n^{-1} \sum_{i=1}^{n} e_{i,\mu,\rho} f(X_i)|\) is sub-gaussian with tail parameter \(\sqrt{V_g/n}\) for each \(j = 1, \ldots, q\), and apply Lemma 67 to obtain \(\mathbb{E}(Z_{n,j}^2|X_1, \ldots, X_n) \leq C_{sg,12}^2(2V_g/n)\), and then \(\mathbb{E}(Z_{n,j}^2) \leq C_{sg,12}^2(2V_g/n)\).

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

\textbf{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

\[ K_f(P_\epsilon, P_\theta; h) \geq f'(e^{-b})\epsilon + f'(e^{P_\theta \epsilon h(x)}) - f^\#(e^{P_\theta h(x)}) - \frac{1}{2} R_{33,b}, \]

where \(R_{33,b} = R_{31,b} \mathbb{V}ar_{P_\theta} h(x) + R_{32,b} \mathbb{V}ar_{P_\theta} h(x)\), \(R_{31,b} = \sup_{|u| \leq b} \frac{d^2}{du^2} \{-f'(e^u)\}\), and \(R_{32,b} = \sup_{|u| \leq b} \frac{d^2}{du^2} f^\#(e^u)\).

(ii) If, in addition, \(\mathbb{E} P_\theta h(x) = 0\) and \(\mathbb{E} P_\theta h(x) \leq 0\), then

\[ K_f(P_\epsilon, P_\theta; h) \geq f'(e^{-b})\epsilon + R_{4,b} \left\{ E_{P_\theta} h(x) - \mathbb{E} P_\theta h(x) \right\} - \frac{1}{2} R_{33,b}, \]

where \(R_{4,b} = \inf_{|u| \leq b} \frac{d^2}{du^2} f^\#(e^u)\).

\textbf{Proof} (i) First, \(K_f(P_\epsilon, P_\theta; h)\) can be bounded as follows:

\[ K_f(P_\epsilon, P_\theta; h) \]
\[ = e \mathbb{E} Q f'(e^{h(x)}) + (1 - e) \mathbb{E} P_\theta f^\#(e^{h(x)}) \]
\[ \geq f'(e^{-b})\epsilon + \mathbb{E} P_\theta f'(e^{h(x)}) - \mathbb{E} P_\theta f^\#(e^{h(x)}) \]
\[ = f'(e^{-b})\epsilon + K_f(P_{\theta^*}, P_{\theta}; h), \]

where the inequality follows because \(f'(e^{h(x)}) \geq f'(e^{-b})\) for \(h(x) \in [-b, b]\) by the convexity of \(f\). Next, consider the function

\[ \kappa(t) = \mathbb{E} P_{\theta^*} f'(e^{E_1 + t\tilde{h}_1(x)}) - \mathbb{E} P_\theta f^\#(e^{E_2 + t\tilde{h}_2(x)}), \]

where \(E_1 = \mathbb{E} P_\theta h(x), E_2 = \mathbb{E} P_\theta h(x), \tilde{h}_1(x) = h(x) - E_1\), and \(\tilde{h}_2(x) = h(x) - E_2\). A Taylor expansion of \(\kappa(1) = K_f(P_{\theta^*}, P_{\theta}; h)\) about \(t = 0\) yields

\[ K_f(P_{\theta^*}, P_{\theta}; h) = f'(e^{E_1}) - f^\#(e^{E_2}) - \frac{1}{2} \kappa''(t), \]

where for some \(t \in [0, 1],\)

\[ \kappa''(t) = -\mathbb{E} P_{\theta^*} \left\{ \tilde{h}_1^2(x) \frac{d^2}{du^2} f'(e^{E_1 + u})|_{u=\tilde{h}_1(x)} \right\} + \mathbb{E} P_\theta \left\{ \tilde{h}_2^2(x) \frac{d^2}{du^2} f^\#(e^{E_2 + u})|_{u=\tilde{h}_2(x)} \right\}. \]
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 $\kappa''(t) \leq R_{33, b}$ by the definition of $R_{33, b}$.

(ii) The inequality from (i) can be rewritten as

$$K_f(P_e, P_\hat{\theta}; h) \geq f'(e^{-b}) + f'(e^{P_{\hat{\theta}}h(x)} - f'(e^{P_{\hat{\theta}}h(x)}) - f'(e^{P_{\hat{\theta}}h(x)}) - \frac{1}{2}R_{33, b}.$$  

If $E_{\hat{\theta}}h(x) = 0$, then

$$K_f(P_e, P_\hat{\theta}; h) \geq f'(e^{-b}) + f'(e^{P_{\hat{\theta}}h(x)} - f'(e^{P_{\hat{\theta}}h(x)}) - \frac{1}{2}R_{33, b}.$$  

Moreover, if $E_{\hat{\theta}}h(x) - E_{\hat{\theta}}h(x) \geq 0$, then

$$f'(e^{P_{\hat{\theta}}h(x)} - f'(e^{P_{\hat{\theta}}h(x)}) \geq R_{4,b} \left\{ E_{\hat{\theta}}h(x) - E_{\hat{\theta}}h(x) \right\},$$  

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 $\gamma \in \Gamma_{\text{rp}1}$ with $\text{pen}_2(\gamma) = b_2$, $E_{\hat{\theta}}h_{\gamma, \hat{\theta}}(x) = 0$, and $E_{\hat{\theta}}h_{\gamma, \hat{\theta}}(x) \leq 0$,

$$K_f(P_{\hat{\theta}}, P_{\hat{\theta}}; h_{\gamma, \hat{\theta}}) \geq R_{4,b} \left\{ E_{\hat{\theta}}h_{\gamma, \hat{\theta}}(x) - E_{\hat{\theta}}h_{\gamma, \hat{\theta}}(x) \right\} + f'(e^{-b}) - 4C_{\hat{\theta}, 12}^2 M_2b_2^2 R_{3,b}^2 - b_2 R_{2,b}^2 \lambda_{22}$$  

where $R_{3,b} = R_{31,b} + R_{32,b}$ as in Lemma 36 and, with $C_{\text{rad}5} = C_{\text{sg}, 12} C_{\text{rad}3}$,

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

depending on the universal constants $C_{\text{sg}, 12}$ and $C_{\text{rad}3}$ in Lemma 67 and Corollary 82.

**Proof** By definition, for any $\gamma \in \Gamma_{\text{rp}1}$, $h_{\gamma}(x)$ can be represented as $h_{\text{rp}1, \beta, c}(x)$ such that $\beta_0 = \gamma_0$ and $\text{pen}_2(\beta) = \sqrt{2}\text{pen}_2(\gamma)$:

$$h_{\text{rp}1, \beta, c}(x) = \beta_0 + \sum_{j=1}^p \beta_1_j \text{ramp}(x_j - c_j)$$  

$$= \beta_0 + \beta_1^T \varphi_{\text{rp},c}(x),$$  

where $c = (c_1, \ldots, c_p)^T$ with $c_j \in \{0, 1\}$, $\beta = (\beta_0, \beta_1^T)^T$ with $\beta_1 = (\beta_1_1, \ldots, \beta_1_p)^T$, and $\varphi_{\text{rp},c}(x) : \mathbb{R}^p \rightarrow [0, 1]^p$ denotes the vector of functions with the $j$th component $\text{ramp}(x_j - c_j)$ for $j = 1, \ldots, p$. Then for any $\gamma \in \Gamma_{\text{rp}1}$ with $\text{pen}_2(\gamma) = b_2$, we have $\beta_0 = \gamma_0$ and $\text{pen}_2(\beta) = \sqrt{2}b_2$ correspondingly, and hence $h_{\gamma}(x) - \gamma_0 = h_{\text{rp}1, \beta, c}(x) - \beta_0 \in [-b_2\sqrt{2p}, b_2\sqrt{2p}]$ by the
Cauchy–Schwartz inequality and the boundedness of the ramp function in $[0, 1]$. Moreover, $h_{\mathrm{trp}, \beta, c}(x)$ can be expressed in the form $\beta_0 + \text{pen}_2(\beta) w^g(x)$, 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 $\text{ramp}(x_j)$ and $\text{ramp}(x_j - 1)$ for $j = 1, \ldots, p$. Parenthetically, at most one of the coefficients in $w$ associated with $\text{ramp}(x_j)$ and $\text{ramp}(x_j - 1)$ is nonzero for each $j$, although this property is not used in the subsequent discussion.

For any $\gamma \in \Gamma_{\mathrm{trp}}$ with $\text{pen}_2(\gamma) = b_2$ and $E_{P_0} h_{\gamma, \hat{\mu}}(x) = 0$, the function $h_{\gamma, \hat{\mu}}(x)$ can be expressed as

$$h_{\gamma, \hat{\mu}}(x) = \beta_1^T \{ \varphi_{\mathrm{trp}, c}(x - \hat{\mu}) - \beta_{01} \},$$

where $\beta_{01} = E_{P_0} \varphi_{\mathrm{trp}, c}(x - \hat{\mu})$. The mean-centered ramp functions in $\varphi_{\mathrm{trp}, c}(x - \hat{\mu}) - \beta_{01}$ 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 $\text{pen}_2(\beta) w^T \{ g(x - \hat{\mu}) - \eta_0 \}$, 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 $\eta_0 = E_{P_0} g(x - \hat{\mu}) \in [0, 1]^q$ by the boundedness of the ramp function in $[0, 1]$.

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

$$K_f(P_n, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}}) \geq K_f(P_n, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}}) - \{ K_f(P_n, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}}) - K_f(P_{\hat{\mu}}, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}}) \}.$$

(119)

For any $\gamma \in \Gamma_{\mathrm{trp}}$ with $\text{pen}_2(\gamma) = b_2$, $E_{P_0} h_{\gamma, \hat{\mu}}(x) = 0$, and $E_{P_0} h_{\gamma, \hat{\mu}}(x) \leq 0$, applying Lemma 46(ii) with $h = h_{\gamma, \hat{\mu}}$ and $b = b_2 = b_2 \sqrt{2p}$ yields

$$K_f(P_n, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}}) \geq f'(e^{-b_2^2}) \epsilon + R_{4, b_2} \left\{ E_{P_0} h(x) - E_{P_0} h(x) \right\} - 2 \left\{ R_{31, b_2} \text{Var}_{P_0} h_{\gamma, \hat{\mu}}(x) + R_{32, b_2} \text{Var}_{P_0} h_{\gamma, \hat{\mu}}(x) \right\}.$$

By Lemma 62(i), $\text{Var}_{P_0} h_{\gamma, \hat{\mu}}(x)$ can be bounded as follows:

$$\text{Var}_{P_0} h_{\gamma, \hat{\mu}}(x) = \text{Var}_{P_0} \beta_1^T \varphi_{\mathrm{trp}, c}(x - \hat{\mu}) \leq \| \beta_1 \|^2 \cdot 2C_{\mathrm{sg}, 12}(\sqrt{2})^2 \| \Sigma \|_{op} = 4\text{pen}_2^2(\beta) C_{\mathrm{sg}, 12}^2 M_2.$$

Similarly, $\text{Var}_{P_0} h_{\gamma, \hat{\mu}}(x)$ can also be bounded by $4\text{pen}_2^2(\beta) C_{\mathrm{sg}, 12}^2 M_2$, because $\| \Sigma \|_{op} \leq M_2$. Hence $K_f(P_n, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}})$ can be bounded as

$$K_f(P_n, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}}) \geq f'(e^{-b_2^2}) \epsilon + R_{4, b_2} \left\{ E_{P_0} h(x) - E_{P_0} h(x) \right\} - 4C_{\mathrm{sg}, 12}^2 M_2 b_2^2 R_{3, b_2},$$

(120)

where $R_{3, b} = R_{31, b} + R_{32, b}$ as in Lemma 36. Moreover, by Lemma 34 with $b = \sqrt{2b_2}$ and $g(x) \in [0, 1]^q$ defined above, it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{\mathrm{trp}}$ with $\gamma_0 = 0$ and $\text{pen}_2(\gamma) = b_2$,

$$\frac{\{ K_f(P_n, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}}) - K_f(P_n, P_{\hat{\mu}}; h_{\gamma, \hat{\mu}}) \}}{b_2 R_{2, b_2} \sqrt{q}} \left\{ C_{\mathrm{sg}, 12} \sqrt{\frac{2qV_0}{n}} + \sqrt{\frac{q \log(\delta^{-1})}{n}} \right\} \leq b_2 R_{2, b_2} \sqrt{2q} \left\{ C_{\mathrm{sg}, 12} \sqrt{\frac{16p}{n}} + \sqrt{\frac{2p \log(\delta^{-1})}{n}} \right\},$$

(121)
where \( V_g = 4C_{\text{rad}3}^2 \) 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 \( \mu = (\mu_1, \ldots, \mu_p)^T \), \( \eta_0 = (\eta_{01}, \ldots, \eta_{0q})^T \), and, as defined in Lemma 34, \( g_{\mu,\eta_0,j}(x) \) is of the form \( \text{ramp}(x_{j1} - \mu_{j1}) - \eta_{0j} \) or \( \text{ramp}(x_{j1} - \mu_{j1} - 1) - \eta_{0j} \). 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 \( \lambda \) and each component in \( \beta \), we obtain that conditionally on \((X_1, \ldots, X_n)\), the random variable \( Z_{n,j} = \sup_{\mu \in \mathbb{R}^p, \eta_0 \in [0,1]^q} |n^{-1} \sum_{i=1}^n \epsilon_i g_{\mu,\eta_0,j}(X_i)| = \sup_{f_j \in \mathcal{G}_j} |n^{-1} \sum_{i=1}^n \epsilon_if_j(X_i)| \) is sub-gaussian with tail parameter \( C_{\text{rad}3}\sqrt{4/n} \) for \( j = 1, \ldots, q \).

Combining the inequalities (119)–(121) leads to the desired result. \( \square \)

**Proposition 48** In the setting of Proposition 24 or Proposition 26, suppose that for \( a \in (0,1/2) \),

\[
D \overset{\text{def}}{=} \sup_{\gamma \in \Gamma_{\text{rp1,pen2}}(\gamma) = \sqrt{1/2}} \left\{ E_{P_\theta} h_{\gamma,\hat{\mu}}(x) - E_{P_\theta} h_{\gamma,\hat{\mu}}(x) \right\} \leq a. \tag{122}
\]

Then we have

\[
\| \hat{\mu} - \mu^* \|_2 \leq S_{4,a} D, \tag{123}
\]

\[
\| \hat{\sigma} - \sigma^* \|_2 \leq S_5 (D + \| \hat{\mu} - \mu^* \|_2 / 2) \leq S_{6,a} D, \tag{124}
\]

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 \( \gamma \in \Gamma_{\text{rp1}} \) with \( \text{pen2}(\gamma) = \sqrt{1/2} \), the function \( h_\gamma(x) \) can be obtained as \( h_{\text{rp1},\beta,c}(x) \) with \( \text{pen2}(\beta) = 1 \). For \( j = 1, \ldots, p \), we restrict \( h_{\gamma,\hat{\mu}}(x) \) in (122) such that \( h_\gamma(x) \) is a ramp function of \( x_{j1} \) in the form \( \pm \text{ramp}(x_{j1} - c) \) for \( c \in \{0,1\} \). Applying Lemma 38 shows that there exists \( h_j^{(1)}(x_j) \) in the form \( \pm \text{ramp}(x_j) \) and \( h_j^{(2)}(x_j) \) in the form \( \pm \text{ramp}(x_j - 1) \) such that

\[
|\hat{\mu}_j - \mu^*_j| \leq S_{4,a} \left\{ E_{P_\theta} h_j^{(1)}(x_j - \hat{\mu}_j) - E_{P_\theta} h_j^{(1)}(x_j - \hat{\mu}_j) \right\}, \tag{125}
\]

\[
|\hat{\sigma}_j - \sigma^*_j| \leq S_5 \left\{ E_{P_\theta} h_j^{(2)}(x_j - \hat{\mu}_j) - E_{P_\theta} h_j^{(2)}(x_j - \hat{\mu}_j) \right\} + S_5 |\hat{\mu}_j - \mu^*_j| / 2. \tag{126}
\]

From (125), we have that for any \( L_2 \) unit vector \( w = (w_1, \ldots, w_p)^T \),

\[
\sum_{j=1}^p |w_j(\hat{\mu}_j - \mu_j)| \leq S_{4,a} \sum_{j=1}^p |w_j| \left\{ E_{P_\theta} h_j^{(1)}(x_j - \hat{\mu}_j) - E_{P_\theta} h_j^{(1)}(x_j - \hat{\mu}_j) \right\}
\]

\[
= S_{4,a} \left\{ E_{P_\theta} h^{(1)}(x - \hat{\mu}) - E_{P_\theta} h^{(1)}(x - \hat{\mu}) \right\},
\]

where \( h^{(1)}(x) = \sum_{j=1}^p |w_j|h_j^{(1)}(x_j) \). In fact, \( h^{(1)}(x) \) can be expressed as \( h_{\text{rp1},\beta,c}(x) \) such that \( c = (0, \ldots, 0)^T \) and each component in \( \beta \) is either \( |w_j| \) or \( -|w_j| \) for \( j = 1, \ldots, p \), which implies that \( \text{pen2}(\beta) = \|w\|_2 = 1 \). Hence by the definition of \( D \), we obtain (123).
Similarly, from (126), we have that for any $L_2$ unit vector $w = (w_1, \ldots, w_p)^T$,
\[
\sum_{j=1}^{p} |w_j(\hat{\sigma}_j - \sigma_j)| \leq S_0 \sum_{j=1}^{p} |w_j| \left\{ E_{P_{\theta}} h_j^{(2)}(x_j - \hat{\mu}_j) - E_{P_{\theta}} h_j^{(2)}(x_j - \hat{\mu}_j) \right\} + S_0 \sum_{j=1}^{p} |w_j(\hat{\mu}_j - \mu_*^j)|/2
\]
and
\[
= S_0 \left\{ E_{P_{\theta}} h^{(2)}(x - \hat{\mu}) - E_{P_{\theta}} h^{(2)}(x - \hat{\mu}) \right\} + S_0 w^T(\hat{\mu} - \mu^*),
\]
where $h^{(2)}(x) = \sum_{j=1}^{p} |w_j|h_j^{(2)}(x_j)$, which can be expressed in the form $h_{\mathrm{rp},\beta,c}(x)$ with $c = (1, \ldots, 1)^T$ and $\pen_2(\beta) = \|w\|_2 = 1$. Hence by the definition of $D$, we obtain (124).

**Proposition 49** Let $b_3 > 0$ be fixed and $b_3^j = b_3 \sqrt{p(p - 1)}$. In the setting of Proposition 24, it holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\pen_2(\gamma) = b_3$, $E_{P_{\theta}} h_{\gamma,\hat{\mu}}(x) \leq 0$,
\[
K_f(P_n, P_{\hat{\mu}}; h_{\gamma,\hat{\mu}})
\geq R_{4,2b_3^j} \left\{ E_{P_{\theta}} h(x) - E_{P_{\theta}} h(x) \right\} + f'(e^{-2b_3^j}) e - (80C_{sg,12}^2 M_2) p b_3^3 R_{3,2b_3^j} - \sqrt{p} b_3 R_{2,b_3^j} \lambda_{32},
\]
where $R_{3,b} = R_{31,b} + R_{32,b}$ as in Lemma 36 and, with $C_{\mathrm{rad}5} = C_{sg,12} C_{\mathrm{rad}3}$,
\[
\lambda_{32} = C_{\mathrm{rad}5} \sqrt{\frac{12(p - 1)}{n}} + \sqrt{\frac{(p - 1) \log(\delta^{-1})}{n}},
\]
depending on the universal constants $C_{sg,12}$ and $C_{\mathrm{rad}3}$ in Lemma 67 and Corollary 82.

**Proof** By definition, for any $\gamma \in \Gamma_{\mathrm{rp}}$, $h_{\gamma}(x)$ can be represented as $h_{\mathrm{rp},\beta}(x)$ such that $\beta_0 = \gamma_0$ and $\pen_2(\beta) = 2\pen_2(\gamma)$, where
\[
h_{\mathrm{rp},\beta}(x) = \beta_0 + \sum_{1 \leq i \neq j \leq p} \beta_{2,ij} \mathrm{ramp}(x_i) \mathrm{ramp}(x_j)
\]
and
\[
\beta = (\beta_0, \beta_2^T) \quad \text{with} \quad \beta_2 = (\beta_{2,ij} : 1 \leq i \neq j \leq p)^T, \quad \text{and} \quad \varphi_{\mathrm{rp}}(x) : \mathbb{R}^p \to [0, 1]^p \quad \text{denotes the vector of functions with the $j$th component $\mathrm{ramp}(x_j)$ for $j = 1, \ldots, p$.}
\]

Then for any $\gamma \in \Gamma_{\mathrm{rp}}$ with $\pen_2(\gamma) = b_3$, we have $\beta_0 = 0$ and $\pen_2(\beta) = 2b_3$ correspondingly, and hence $h_{\gamma}(x) - \gamma_0 = h_{\mathrm{rp},\beta}(x) - \beta_0 \in [-2b_3 \sqrt{p(p - 1)}, 2b_3 \sqrt{p(p - 1)}]$, 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, $h_{\mathrm{rp},\beta}(x)$ can be expressed in the form $\beta_0 + \pen_2(\beta) w^T g(x)$, 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 $w$. 

82
For any \( \gamma \in \Gamma_{\tau_2} \) with \( \text{pen}_2(\gamma) = b_3 \) and \( E_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) = 0 \), the function \( h_{\gamma,\hat{\mu}}(x) \) can be expressed as

\[
h_{\gamma,\hat{\mu}}(x) = \beta_2^T \{ \varphi_{\tau_2}(x - \hat{\mu}) - \beta_0 \},
\]
where \( \beta_0 = E_{P_\theta^*} \varphi_{\tau_2}(x - \hat{\mu}) \). The mean-centered ramp functions in \( \varphi_{\tau_2}(x - \hat{\mu}) - \beta_0 \) are bounded between \([-1, 1]\), and hence \( h_{\gamma,\hat{\mu}}(x) \in [-b_32p, b_32p] \) similarly as above. Moreover, such \( h_{\gamma,\hat{\mu}}(x) \) can be expressed in the form \( \text{pen}_2(\beta)w^T\{g(x - \hat{\mu}) - \eta_0\} \), where \( w \in \mathbb{R}^q \) is an \( L_2 \) unit vector, \( g(x) : \mathbb{R}^p \rightarrow [0, 1]^q \) is defined as above, and \( \eta_0 = E_{P_\theta^*}g(x - \hat{\mu}) \in [0, 1]^q \) 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

\[
K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \geq K_f(P, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - |K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_f(P, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})|.
\]

(127)

For any \( \gamma \in \Gamma_{\tau_2} \) with \( \text{pen}_2(\gamma) = b_3 \), \( E_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) = 0 \), and \( E_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) \leq 0 \), applying Lemma 46(ii) with \( h = h_{\gamma,\hat{\mu}} \) and \( b = 2b_3^T = 2b_3 \sqrt{p(p - 1)} \) yields

\[
K_f(P, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) 
\geq f'(e^{-2b_3^T}) + R_{4,2b_3^T} \left\{ E_{P_\theta^*} h(x) - E_{P_\theta^*} h(x) \right\} - \frac{1}{2} \left\{ R_{31,2b_3^T} \text{Var}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) + R_{32,2b_3^T} \text{Var}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) \right\}.
\]

By Lemma 63(ii), \( \text{Var}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) \) can bounded as follows:

\[
\text{Var}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) 
= \text{Var}_{P_\theta^*} \beta_2^T \text{vec}(\varphi_{\tau_2}(x - \hat{\mu}) \otimes \varphi_{\tau_2}(x - \hat{\mu})) 
\leq \|\beta_2\|^2 \frac{20C^2_{s_{g,12}}(\sqrt{2})^p\|\Sigma^*\|_{op}}{2} 
\leq 40\text{pen}_2(\beta)C^2_{s_{g,12}}pM_2.
\]

Similarly, \( \text{Var}_{P_\theta^*} h_{\gamma,\hat{\mu}}(x) \) can also be bounded by \( 40\text{pen}_2(\beta)C^2_{s_{g,12}}pM_2 \), because \( \|\Sigma\|_{op} \leq M_2 \). Hence, with \( \text{pen}_2(\beta) = 2b_3 \), \( K_f(P, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) \) can be bounded as

\[
K_f(P, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) 
\geq f'(e^{-2b_3^T}) + R_{4,2b_3^T} \left\{ E_{P_\theta^*} h(x) - E_{P_\theta^*} h(x) \right\} - 80C^2_{s_{g,12}}pM_2b_3^T R_{3,2b_3^T},
\]

(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 \( \gamma \in \Gamma_{\tau_2} \) with \( \gamma_0 = 0 \) and \( \text{pen}_2(\gamma) = b_3 \),

\[
|K_f(P_n, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}}) - K_f(P, P_{\hat{\theta}}; h_{\gamma,\hat{\mu}})| 
\leq b_3 R_{2,b_3^T} \left\{ C_{s_{g,12}} \sqrt{\frac{2qV_g}{n} + \sqrt{\frac{q\log(\delta^{-1})}{n}}} \right\} 
= b_3 R_{2,b_3^T} \left\{ C_{s_{g,12}}C_{\text{rad}} \sqrt{\frac{12p(p - 1)}{n} + \sqrt{\frac{p(p - 1)\log(\delta^{-1})}{n}}} \right\},
\]

(129)
where $V_g = 6C_{\text{rad3}}^2$ is determined in Lemma 34 as follows. For $j = 1, \ldots, q$, consider the function class $G_j = \{g_{\mu, \eta_0,j} : \mu \in \mathbb{R}^p, \eta_0 \in [0, 1]^q\}$, where $\mu = (\mu_1, \ldots, \mu_p)^T$, $\eta_0 = (\eta_01, \ldots, \eta_0q)^T$, and, as defined in Lemma 34, $g_{\mu, \eta_0,j}(x)$ is of the form $\text{ramp}(x_{j_1} - \mu_{j_1}) \text{ramp}(x_{j_2} - \mu_{j_2}) - \eta_{0j}$ for $1 \leq j_1 \neq j_2 \leq p$. 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 $Z_{n,j} = \sup_{\mu \in \mathbb{R}^p, \eta_0 \in [0, 1]^q} |n^{-1} \sum_{i=1}^n \epsilon_i g_{\mu, \eta_0,j}(X_i)| = \sup_{f_j \in G_j} |n^{-1} \sum_{i=1}^n \epsilon_i f_j(X_i)|$ is subgaussian with tail parameter $C_{\text{rad3}}^2 \sqrt{6/n}$ 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

$$D = \sup_{\gamma \in \Gamma_{\text{rp2}}, \text{pen}_2(\gamma) = 1/2} \left\{ \mathbb{E}_{P_\theta^*} h_{\gamma, \hat{\mu}}(x) - \mathbb{E}_{P_\theta} h_{\gamma, \hat{\mu}}(x) \right\},$$

Then we have

$$\| \hat{\Sigma} - \Sigma^* \|_F \leq 2M_2^{1/2} \sqrt{p} \| \hat{\sigma} - \sigma^* \|_2 + S_7(\sqrt{2p \Delta_{\hat{\mu}, \hat{\sigma}} + D),$$

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 $\gamma \in \Gamma_{\text{rp2}}$ with $\text{pen}_2(\gamma) = 1/2$, the function $h_\gamma(x) \in \mathcal{H}_{\text{rp2}}$ can be obtained as $h_{\text{rp2}, \beta}(x)$ with $\text{pen}_2(\beta) = 1$. First, we handle the effect of different means and standard deviations between $P_\theta^*$ and $P_\theta$ in $D$. Denote

$$D^\dagger = \sup_{\gamma \in \Gamma_{\text{rp2}}, \text{pen}_2(\gamma) = 1/2} \left\{ \mathbb{E}_{P_\theta^*} h_{\gamma, \hat{\mu}}(x) - \mathbb{E}_{P_\theta} h_{\gamma, \hat{\mu}}(x) \right\},$$

where $\theta^\dagger = (\hat{\mu}, \text{diag}(\hat{\sigma}) \Sigma_0^* \text{diag}(\hat{\sigma}))$ and $\Sigma_0^*$ is defined as the correlation matrix such that $\Sigma^* = \text{diag}(\sigma^*) \Sigma_0^* \text{diag}(\sigma^*)$. Then $D^\dagger$ can be related to $D$ as follows:

$$D^\dagger \leq D + \sqrt{2p \Delta_{\hat{\mu}, \hat{\sigma}},}$$

(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 $\varphi_{\text{rp}}(x)$, which is $(1/2)$-Lipschitz and componentwise bounded in $[0, 1]$, we have that for any $\gamma \in \Gamma_{\text{rp2}}$ with $\text{pen}_2(\gamma) = 1/2$,

$$\left| \mathbb{E}_{P_\theta^*} h_{\gamma, \hat{\mu}}(x) - \mathbb{E}_{P_\theta} h_{\gamma, \hat{\mu}}(x) \right| \leq \sqrt{2p \Delta_{\hat{\mu}, \hat{\sigma}}.$$

For each pair $1 \leq i \neq j \leq p$, we restrict $h_{\gamma, \hat{\mu}}(x)$ such that $h_{\gamma}(x)$ is ramp$(x_i)$ramp$(x_j)$ or $-\text{ramp}(x_i)\text{ramp}(x_j)$. Applying Lemma 40 shows that there exists $r_{ij} \in \{-1, 1\}$ such that

$$|\hat{\rho}_{ij} - \rho_{ij}^\dagger| \hat{\sigma}_{ij} \leq S_7 r_{ij} \left\{ \mathbb{E}_{P_\theta^*} h_{ij}(x - \hat{\mu}) - \mathbb{E}_{P_\theta} h_{ij}(x - \hat{\mu}) \right\},$$

84
where \( h_{ij}(x) = \text{ramp}(x_i) \text{ramp}(x_j) \). By the triangle inequality, we have
\[
|\hat{\rho}_{ij}\hat{\sigma}_i \hat{\sigma}_j - \rho_{ij}^* \sigma_i^* \sigma_j^*| \\
\leq |\hat{\sigma}_i |\hat{\sigma}_j - \sigma_j^*| + \sigma_i^*|\hat{\sigma}_i - \sigma_i^*| + |\hat{\rho}_{ij} - \rho_{ij}^*| |\hat{\sigma}_j| \\
\leq M_2^{1/2}|\hat{\sigma}_i - \sigma_i^*| + M_2^{1/2}|\hat{\sigma}_j - \sigma_j^*| + S_T r_{ij} \left\{ E_{P_{\hat{\sigma}}^i} h_{ij}(x - \hat{\mu}) - E_{P_{\hat{\sigma}}} h_{ij}(x - \hat{\mu}) \right\}.
\]
In addition, we have \(|\hat{\sigma}_i^2 - \sigma_i^*|^2 = |(\hat{\sigma}_i + \sigma_i^*)(\hat{\sigma}_i - \sigma_i^*)| \leq 2M_2^{1/2}|\hat{\sigma}_i - \sigma_i^*|\). Then for any \( L_2 \) unit vector \( w = (w_{ij} : 1 \leq i, j \leq p)^T \in \mathbb{R}^{p \times p}, \)
\[
\sum_{i=1}^p |w_{ii}(\hat{\sigma}_i^2 - \sigma_i^*|^2) + \sum_{1 \leq i \neq j \leq p} |w_{ij}(\hat{\rho}_{ij}\hat{\sigma}_i \hat{\sigma}_j - \rho_{ij}^* \sigma_i^* \sigma_j^*)| \\
\leq 2M_2^{1/2} \sum_{i=1}^p |w_{ii}| |\hat{\sigma}_i - \sigma_i^*| + M_2^{1/2} \sum_{1 \leq i \neq j \leq p} |w_{ij}| (|\hat{\sigma}_i - \sigma_i^*| + |\hat{\sigma}_j - \sigma_j^*|) \\
+ S_T \sum_{1 \leq i \neq j \leq p} |w_{ij}| r_{ij} \left\{ E_{P_{\hat{\sigma}}^i} h_{ij}(x - \hat{\mu}) - E_{P_{\hat{\sigma}}} h_{ij}(x - \hat{\mu}) \right\} \\
= M_2^{1/2} \sum_{1 \leq i \leq p} |w_{ij}| (|\hat{\sigma}_i - \sigma_i^*| + |\hat{\sigma}_j - \sigma_j^*|) + S_T \left\{ E_{P_{\hat{\sigma}}^i} h(x - \hat{\mu}) - E_{P_{\hat{\sigma}}} h(x - \hat{\mu}) \right\},
\]
where \( h(x) = \sum_{1 \leq i \neq j \leq p} |w_{ij}| r_{ij} h_{ij}(x) \). The function \( h(x) \) can be expressed as \( h_{\beta_{2,p}}(x) \) such that \( \beta_{2,ii} = 0 \) for \( i = 1, \ldots, p \) and \( \beta_{2,ij} = |w_{ij}| r_{ij} \) for \( 1 \leq i \neq j \leq p \), and hence \( \text{pen}_2(\beta) \leq \|w\|_2 = 1 \). By the definition of \( D^1 \), we have \( E_{P_{\beta}}^i h(x - \hat{\mu}) - E_{P_{\hat{\sigma}}} h(x - \hat{\mu}) \leq D^1 \). Moreover, by the Cauchy–Schwarz inequality, \( \sum_{1 \leq i \leq p} |w_{ii}| |\hat{\sigma}_i - \sigma_i^*| \leq \sqrt{p} \|\hat{\sigma} - \sigma^*\|_2 \). Substituting these inequalities into the preceding display shows that
\[
\sum_{i=1}^p |w_{ii}(\hat{\sigma}_i^2 - \sigma_i^*|^2) + \sum_{1 \leq i \neq j \leq p} |w_{ij}(\hat{\rho}_{ij}\hat{\sigma}_i \hat{\sigma}_j - \rho_{ij}^* \sigma_i^* \sigma_j^*)| \\
\leq 2M_2^{1/2} \sqrt{p} \|\hat{\sigma} - \sigma^*\|_2 + S_T D^1.
\]
Combining (130) and (131) yields the desired result.

\[\Box\]

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
\[
\lambda_{11} = \sqrt{\frac{2 \log(5p) + \log(\delta^{-1})}{n}} + \frac{2 \log(5p) + \log(\delta^{-1})}{n},
\]
\[
\lambda_{12} = 2C_{rad4} \sqrt{\frac{\log(2p(p + 1))}{n}} + \sqrt{\frac{2 \log(\delta^{-1})}{n}},
\]
where \( C_{rad4} = C_{sg6}C_{rad3} \), depending on universal constants \( C_{sg6} \) and \( C_{rad4} \) in Lemmas 70 and Corollary 82 in Appendix E. Denote
\[
\text{Err}_{h1}(n, p, \delta, \epsilon) = 3\epsilon + 2\sqrt{\epsilon/(n\delta)} + \lambda_{12} + \lambda_1.
\]
Proposition 21 (restated) Assume that \( \| \Sigma_* \|_{\text{max}} \leq M_1 \) and \( \epsilon \leq 1/5 \). Let \( \hat{\theta} = (\hat{\mu}, \hat{\Sigma}) \) be a solution to (20) with \( \lambda_1 \geq C_{\text{sp13}} M_{11} \lambda_{11} \) 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 \( C_{\text{sp11}} \) and \( C_{\text{sp12}} \) in Lemma 30 in Appendix C. If \( \sqrt{\epsilon(1-\epsilon)/(n\delta)} \leq 1/5 \) and \( \text{Err}_{h1}(n, p, \delta, \epsilon) \leq a \) for a constant \( a \in (0, 1/2) \), then the following holds with probability at least \( 1 - 7\delta \) uniformly over contamination distribution \( Q \),

\[
\| \hat{\mu} - \mu^* \|_{\infty} \leq S_{4,a} \text{Err}_{h1}(n, p, \delta, \epsilon),
\]

\[
\| \hat{\Sigma} - \Sigma^* \|_{\text{max}} \leq S_{8,a} \text{Err}_{h1}(n, p, \delta, \epsilon),
\]

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

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) \( \lambda_1 \geq C_1 \left( \sqrt{\log p/n} + \sqrt{\log (1/\delta)/n} \right) \),

(ii) \( \epsilon + \sqrt{\epsilon/(n\delta)} + \lambda_1 \leq C_2 \),

imply the conditions

(iii) \( \lambda_1 \geq C_{\text{sp13}} M_{11} \lambda_{11} \),

(iv) \( \text{Err}_{h1}(n, p, \delta, \epsilon) \leq a \),

(v) \( \epsilon \leq 1/5 \) and \( \sqrt{\epsilon(1-\epsilon)/(n\delta)} \leq 1/5 \).

In fact, condition (v) follows directly from condition (ii) because \( C_2 \leq 1/5 \). For conditions (iii) and (iv), we first show that \( \lambda_{11} \) and \( \lambda_{12} \) can be upper bounded as follows:

\[
\lambda_{11} \leq \sqrt{\frac{2 \log p + 3 \log (\delta^{-1})}{n}} + \frac{2 \log p + 3 \log (\delta^{-1})}{n}, \tag{132}
\]

\[
\leq 2 \sqrt{\frac{2 \log p + 3 \log (\delta^{-1})}{n}} \leq 2 \sqrt{\frac{2 \log p}{n}} + 2 \sqrt{3 \log (\delta^{-1})}, \tag{133}
\]

86
and
\[ \lambda_{12} \leq 2C_{\text{rad}4} \sqrt{\frac{2 \log 2 + 2 \log p}{n}} + \sqrt{\frac{2 \log(\delta^{-1})}{n}} \quad (134) \]
\[ \leq 2C_{\text{rad}4} \sqrt{\frac{2 \log p}{n}} + (2C_{\text{rad}4} + 1) \sqrt{\frac{2 \log(\delta^{-1})}{n}}. \quad (135) \]

Lines (132) and (135) hold because \( \log(1/\delta) \geq \log(5) \) for \( \delta \in (0, 1/7) \). Line (133) holds because
\[ \sqrt{\frac{2 \log p + 3 \log(\delta^{-1})}{n}} \leq 1 \]
and hence the linear term in \( \lambda_{11} \) is upper bounded by the square root term. To see this, by conditions (i) and (ii) we have
\[ \sqrt{\frac{2 \log p + 3 \log(\delta^{-1})}{n}} = \sqrt{\frac{3 \log p}{n}} + \sqrt{\frac{3 \log(\delta^{-1})}{n}} \leq \frac{\sqrt{3} \lambda_1}{C_1} \leq \frac{\sqrt{3} C_2}{C_1} \leq 1. \]

Line (134) holds because \( \log(2p(p+1)) \leq 2 \log 2 + 2 \log p \) for \( p \geq 1 \). With the above upper bounds for \( \lambda_{11} \) and \( \lambda_{12} \), we show that condition (i) implies condition (iii) as follows:
\[ \lambda_1 \geq C_1 \left( \sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \]
\[ = 2\sqrt{3} C_{\text{sp}13} M_{11} \left( \sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \geq C_{\text{sp}13} M_{11} \lambda_{11}, \]

and condition (ii) implies condition (iv) as follows:
\[ \text{Err}_{h_1}(n, p, \delta, \epsilon) = 3\epsilon + 2\sqrt{\epsilon/(n\delta)} + \lambda_{12} + \lambda_1 \]
\[ \leq 3 \left( \epsilon + \sqrt{\frac{\epsilon}{n\delta}} \right) + \left( \frac{\sqrt{2} C_1}{4C_{\text{rad}4} + 2} + 1 \right) \lambda_1 \]
\[ \leq \left( 3 \vee \left( \frac{\sqrt{2} C_1}{4C_{\text{rad}4} + 2} + 1 \right) \right) \left( \epsilon + \sqrt{\frac{\epsilon}{n\delta}} + \lambda_1 \right) \]
\[ \leq \left( 3 \vee \left( \frac{\sqrt{2} C_1}{4C_{\text{rad}4} + 2} + 1 \right) \right) C_2 \leq a. \]

Therefore, Proposition 21 implies Theorem 15 with constant
\[ C = S_{8,a} \left( 3 \vee \left( \frac{\sqrt{2} C_1}{4C_{\text{rad}4} + 2} + 1 \right) \right). \]

**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
\[ K_{\text{HG}}(P, P_{\theta^*}; h) \leq 2\epsilon. \]
(ii) For any function \( h : \mathbb{R}^p \to \mathbb{R} \), we have
\[ K_{\text{HG}}(P_n, P_{\theta^*}; h) \leq 2\hat{\epsilon} + |E_{P_{\theta^*, n}} h(x) - E_{P_{\theta^*}} h(x)|, \quad (136) \]
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

\[
K_{HG}(P_e, P_{\theta^*}; h) = \epsilon E_Q \min(1, h(x)) + (1 - \epsilon) E_{P_{\theta^*}} \min(1, h(x)) + E_{P_{\theta^*}} \min(1, -h(x)) \\
\leq \epsilon + (1 - \epsilon) E_{P_{\theta^*}} \min(1, h(x)) + E_{P_{\theta^*}} \min(1, -h(x)) \quad (137) \\
\leq 2\epsilon + (1 - \epsilon) \{ E_{P_{\theta^*}} \min(1, h(x)) + E_{P_{\theta^*}} \min(1, -h(x)) \} \quad (138) \\
\leq 2\epsilon. \quad (139)
\]

Inequalities (137) and (138) hold because \( \min(1, u) \vee \min(1, -u) \leq 1 \) for all \( u \in \mathbb{R} \). Inequality (139) holds because \( \min(1, u) + \min(1, -u) \leq 0 \) 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

\[
K_{HG}(P_n, P_{\theta^*}; h) = \frac{1}{n} \sum_{i=1}^{n} R_i \min(1, h(X_i)) + \frac{1}{n} \sum_{i=1}^{n} (1 - R_i) \min(1, h(X_i)) + E_{P_{\theta^*}} \min(1, -h(X_i)) \\
\leq \hat{\epsilon} + (1 - \hat{\epsilon}) E_{P_{\theta^*},n} \min(1, h(X_i)) + E_{P_{\theta^*}} \min(1, -h(X_i)) \quad (140) \\
\leq \hat{\epsilon} + (1 - \hat{\epsilon}) \min(1, E_{P_{\theta^*},n} h(X_i)) + \min(1, -E_{P_{\theta^*}} h(X_i)) \quad (141) \\
\leq 2\hat{\epsilon} + (1 - \hat{\epsilon}) \{ \min(1, E_{P_{\theta^*},n} h(X_i)) + \min(1, -E_{P_{\theta^*}} h(X_i)) \} \quad (142) \\
\leq 2\hat{\epsilon} + 0 + | \min(1, -E_{P_{\theta^*},n} h(x)) - \min(1, -E_{P_{\theta^*}} h(x)) | \quad (143) \\
\leq 2\hat{\epsilon} + 0 + | E_{P_{\theta^*},n} h(x) - E_{P_{\theta^*}} h(x) | \quad (144)
\]

Lines (140) and (142) hold because \( \min(1, u) \vee \min(1, -u) \leq 1 \) 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 \( \min(1, u) + \min(1, -u) \leq 0 \) 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 \( \gamma \in \Gamma \),

\[
K_{HG}(P_n, P_{\theta^*}; h_{\gamma, \theta^*}) \leq 2(\epsilon + \sqrt{\epsilon/(n\delta)}) + \text{pen}_1(\gamma) C_{\text{sp}13} M_{11} \lambda_{11},
\]

where \( C_{\text{sp}13} = (5/3)(C_{\text{sp}11} \vee C_{\text{sp}12}) \) with \( C_{\text{sp}11} \) and \( C_{\text{sp}12} \) 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 \( \Omega_1 \) and \( \Omega_2 \). In the event \( \Omega_1 \) we have \( |\hat{\epsilon} - \epsilon| \leq 1/5 \) by the assumption \( \sqrt{\epsilon(1 - \epsilon)/(n\delta)} \leq 1/5 \).
and hence \( \hat{e} \leq 2/5 \) by the assumption \( \epsilon \leq 1/5 \). Thus, by Lemma 51 with \( \epsilon_1 = 2/5 \), it holds in the event \( \Omega_1 \) that for any \( \gamma \in \Gamma \),

\[
K_{HG}(P_n, P_\theta^*; h_{\gamma, \mu^*}) \\
\leq 2\hat{e} + \left| E_{P_\theta^*, n} h_{\gamma, \mu^*}(x) - E_{P_\theta, h_{\gamma, \mu^*}(x)} \right| \\
\leq 2(\epsilon + \sqrt{\epsilon/(n\delta)}) + \left| E_{P_\theta^*, n} h_{\gamma}(x - \mu^*) - E_{P_\theta(0, \Sigma^*)} h_{\gamma}(x) \right|. \tag{145}
\]

The last step (145) uses the fact that \( E_{P_\theta, h_{\gamma, \mu^*}(x)} = E_{P(0, \Sigma^*)} h_{\gamma}(x) \) and \( E_{P_\theta^*, n} h_{\gamma, \mu^*}(x) = E_{P_\theta, h_{\gamma}(x - \mu^*)} \), by the definition \( h_{\gamma, \mu^*}(x) = h_{\gamma}(x - \mu^*) \).

Next, as shown in Proposition 33, it holds in the event \( \Omega_2 \) while conditionally on \( \Omega_1 \) that for any \( \gamma = (\gamma_0, \gamma_1^T, \gamma_2^T)^T \in \Gamma \),

\[
\left| E_{P_\theta^*, n} h_{\gamma}(x - \mu^*) - E_{P(0, \Sigma^*)} h_{\gamma}(x) \right| \\
\leq \text{pen}_1(\gamma)(5/3)(C_{sp11} \lor C_{sp12}) M_1 \lambda_1, \tag{146}
\]

where \( h_{\gamma}(x) = \gamma_0 + \gamma_1^T \varphi(x) + \gamma_2^T (\varphi(x) \otimes \varphi(x)) \) and \( \text{pen}_1(\gamma) = \|\gamma_1\|_1 + \|\gamma_2\|_1 \). 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 \( \gamma \in \Gamma_{rp} \) with \( \gamma_0 = 0 \) and \( \text{pen}_1(\gamma) = 1 \),

\[
K_{HG}(P_n, P_{\hat{\beta}}; h_{\gamma, \hat{\mu}}) \\
\geq \left\{ E_{P_\theta, h_{\gamma, \hat{\mu}}(x)} - E_{P_\theta, h_{\gamma, \hat{\mu}}(x)} \right\} - \epsilon - \lambda_{12}
\]

where, with \( C_{rad4} = C_{sg6}C_{rad3} \) is the same constant in Proposition 37,

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

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

\[\]  

**Proof** By definition, for any \( \gamma \in \Gamma_{rp} \), \( h_{\gamma}(x) \) can be represented as \( h_{rp, \beta, c}(x) \) such that \( \beta_0 = \gamma_0 \) and \( \text{pen}(\beta) = \text{pen}(\gamma) \) in the same way as in Proposition 37. Then for any \( \gamma \in \Gamma_{rp} \) with \( \gamma_0 = 0 \) and \( \text{pen}_1(\gamma) = 1 \), we have \( \beta_0 = 0 \) and \( \text{pen}_1(\gamma) = 1 \) correspondingly, and hence \( h_{\gamma}(x) = h_{rp, \beta, c}(x) \in [0, 1] \) by the boundedness of the ramp function in \( [0, 1] \). Moreover, \( h_{rp, \beta, c}(x) \) with \( \beta_0 = 0 \) and \( \text{pen}_1(\gamma) = 1 \) can be expressed in the form \( w^T g(x) \), 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 \( \text{ramp}(x_j) \) and \( \text{ramp}(x_j - 1) \) for \( j = 1, \ldots, p \), and \( \text{ramp}(x_i) \text{ramp}(x_j) \) for \( 1 \leq i \neq j \leq p \). For symmetry, \( \text{ramp}(x_i) \text{ramp}(x_j) \) and \( \text{ramp}(x_j) \text{ramp}(x_i) \) are included as two distinct components in \( g \), and the corresponding coefficients are identical to each other in \( w \).
Next, \( K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) \) can be bounded as
\[
K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) \\
\geq K_{HG}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - \{K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - K_{HG}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}})\}. \tag{147}
\]
For any \( \gamma \in \Gamma_{\text{tr}} \) with \( \gamma_0 = 0 \) and \( \text{pen}_1(\gamma) = 1 \), 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 \( K_{HG}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) \) in (147) reduces to a moment matching term and can be lower bounded as follows:
\[
K_{HG}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) \\
= E_{P_\epsilon} \min(h_{\gamma, \hat{\mu}}(x), 1) + E_{P_{\hat{\theta}}} \min(-h_{\gamma, \hat{\mu}}(x), 1) \\
= E_{P_\epsilon} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \\
= E_{P_\epsilon} h_{\gamma, \hat{\mu}}(x) + (1 - \epsilon) E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \\
\geq -\epsilon + \left\{ E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \right\}.
\]
Similarly, the two other hinge terms in (147) also reduce to moment matching terms:
\[
\{K_{HG}(P_n, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}}) - K_{HG}(P_\epsilon, P_{\hat{\theta}}; h_{\gamma, \hat{\mu}})\} \\
= \left\{ E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \right\} - \left\{ E_{P_\epsilon} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \right\} \\
= E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x).
\]
We apply Lemma 34 with \( b = 1, g(x) \in [0, 1]^q \) defined above, and \( f'(e^u) \) and \( f'(e^u) \) replaced by the identity function in \( u \). It holds with probability at least \( 1 - 2\delta \) that for any \( \gamma \in \Gamma_{\text{tr}} \) with \( \gamma_0 = 0 \) and \( \text{pen}_1(\gamma) = b_1 \),
\[
E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\hat{\theta}}} h_{\gamma, \hat{\mu}}(x) \\
\leq C_{\text{sg6}} \sqrt{\frac{V_g \log(2q)}{n}} + \sqrt{\frac{2 \log(\delta^{-1})}{n}} \\
= C_{\text{sg6}} C_{\text{rad3}} \sqrt{\frac{4 \log(2p)}{n}} + \sqrt{\frac{2 \log(\delta^{-1})}{n}},
\]
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 \( \gamma = (\gamma_0, \gamma_1, \gamma_2)^T \in \text{Gamma}, \)
\[
K_{HG}(P_n, P_{\theta^*}; h_{\gamma, \theta^*}) \leq 2(\epsilon + \sqrt{\epsilon/(n\delta)}) + \text{pen}_2(\gamma_1)(5/3)C_{\text{sp21}} M_2^{1/2} R_1 \lambda_{21} \\
+ \text{pen}_2(\gamma_2)(25\sqrt{5}/3)C_{\text{sp22}} M_2 R_1 \sqrt{p} \lambda_{31},
\]

90
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 \( \Omega_1 \) and \( \Omega_2 \). In the event \( \Omega_1 \) we have \( |\epsilon - \hat{\epsilon}| \leq 1/5 \) by the assumption \( \sqrt{\epsilon(1 - \epsilon)/(n\delta)} \leq 1/5 \) and hence \( \epsilon \leq 2/5 \) by the assumption \( \epsilon \leq 1/5 \). By Lemma 51 with \( \epsilon_1 = 2/5 \), it holds that in the event \( \Omega_1 \) for any \( \gamma \in \Gamma \),

\[
K_{HG}(P_n, P_\theta^*; h_{\gamma, \mu^*}) \\
\leq 2\hat{\epsilon} + \left| E_{P_{\theta^*,n}} h_{\gamma, \mu^*}(x) - E_{P_\theta} h_{\gamma, \mu^*}(x) \right| \\
\leq 2(\epsilon + \sqrt{\epsilon/(n\delta)}) + \left| E_{P_{\theta^*,n}} h_{\gamma}(x - \mu^*) - E_{P_0(\Sigma^*)} h_{\gamma}(x) \right| . \tag{148}
\]

The last step (148) uses the fact that \( E_{P_\theta} h_{\gamma, \mu^*}(x) = E_{P_{0(\Sigma^*)}} h_{\gamma}(x) \) and \( E_{P_{\theta^*,n}} h_{\gamma, \mu^*}(x) = E_{P_{\theta^*,n}} h_{\gamma}(x - \mu^*) \), by the definition \( h_{\gamma, \mu^*}(x) = h_{\gamma}(x - \mu^*) \).

Next, as shown in Proposition 44, it holds in event \( \Omega_2 \) while conditionally on \( \Omega_1 \) that for any \( \gamma = (\gamma_0, \gamma_1, \gamma_2)^T \in \Gamma \),

\[
\left| E_{P_{\theta^*,n}} h_{\gamma}(x - \mu^*) - E_{P_0(\Sigma^*)} h_{\gamma}(x) \right| \\
\leq \text{pen}_2(\gamma_1)(5/3)C_{sp21}M_2^{1/2}\lambda_{21} + \text{pen}_2(\gamma_2)(5/3)C_{sp22}3M_{21}\sqrt{3p}\lambda_{31}, \tag{149}
\]

where \( h_{\gamma}(x) = \gamma_0 + \gamma_1^T \varphi(x) + \gamma_2^T (\varphi(x) \otimes \varphi(x)) \), \( \text{pen}_2(\gamma_1) = \|\gamma_1\|_2 \), and \( \text{pen}_2(\gamma_2) = \|\gamma_2\|_2 \). 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 \( \gamma \in \Gamma \). \( \square \)

**Proposition 55** In the setting of Proposition 26, it holds with probability at least \( 1 - 2\delta \) that for any \( \gamma \in \Gamma_{10} \),

\[
K_{HG}(P_n, P_\theta^*; h_{\gamma, \hat{\mu}}) \\
\geq \left\{ E_{P_\theta} h_{\gamma, \hat{\mu}}(x) - E_{P_\theta^*} h_{\gamma, \hat{\mu}}(x) \right\} - \epsilon - \lambda_{22}(2p)^{-1/2}
\]

where, with \( C_{rad5} = C_{sg, 12}C_{rad3} \), and

\[
\lambda_{22} = C_{rad5} \sqrt{\frac{16p}{n}} + \sqrt{\frac{2p \log(\delta^{-1})}{n}},
\]

depending on the universal constants \( C_{sg, 12} \) and \( C_{rad3} \) in Lemma 67 and Corollary 82.

**Proof** For any \( \gamma \in \Gamma_{10} \subset \Gamma_{rpl} \), because \( \text{pen}_2(\gamma) = (2p)^{-1/2} \) and \( \Gamma_0 = 0 \), we have \( \beta_0 = 0 \) and \( \text{pen}_2(\beta) = p^{-1/2} \). Hence, \( h_{\gamma}(x) = h_{rpl, \beta, c}(x) \in [-1, 1] \) 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 \( E_{P_\theta} h_{\gamma, \hat{\mu}}(x) = 0 \), is also bounded in \([-1, 1]\). Moreover, such \( h_{\gamma, \hat{\mu}}(x) \) can be expressed
in the form \( \text{pen}_2(\beta)w^T \{g(x - \hat{\mu}) - \eta_0\} \), 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 \( \text{ramp}(x_j) \) and \( \text{ramp}(x_j - 1) \) for \( j = 1, \ldots, p \), and \( \eta_0 = E_{P_{\theta'}}g(x - \hat{\mu}) \in [0, 1]^q \).

Next, \( K_{HG}(P_n, P_{\theta'}; h_{\gamma, \hat{\mu}}) \) can be bounded as

\[
K_{HG}(P_n, P_{\theta'}; h_{\gamma, \hat{\mu}}) \geq K_{HG}(P, P_{\theta'}; h_{\gamma, \hat{\mu}}) - \{K_{HG}(P_n, P_{\theta'}; h_{\gamma, \hat{\mu}}) - K_{HG}(P, P_{\theta'}; h_{\gamma, \hat{\mu}})\}.
\]  

(150)

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 \( K_{HG}(P_n, P_{\theta'}; h_{\gamma, \hat{\mu}}) \) reduces to a moment matching term and can be lower bounded as follows:

\[
K_{HG}(P, P_{\theta'}; h_{\gamma, \hat{\mu}})
= E_P \min(h_{\gamma, \hat{\mu}}(x), 1) + E_{P_{\theta'}} \min(-h_{\gamma, \hat{\mu}}(x), 1)
= E_P h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x)
= E_P h_{\gamma, \hat{\mu}}(x) + (1 - e) E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x)
\geq -e + \left\{ E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) \right\}.
\]  

(151)

Similarly, the absolute difference term in (150) can be simplified as follows:

\[
\{K_{HG}(P_n, P_{\theta'}; h_{\gamma, \hat{\mu}}) - K_{HG}(P, P_{\theta'}; h_{\gamma, \hat{\mu}})\}
= \left\{ E_{P_n} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) \right\} - \left\{ E_{P} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) \right\}
= E_{P_n} h_{\gamma, \hat{\mu}}(x) - E_{P} h_{\gamma, \hat{\mu}}(x).
\]

We apply Lemma 45 with \( b = 1, g(x) \in [0, 1]^q \) and \( \eta_0 \) defined above, and \( f'(e^u) \) and \( f'(e^{u'}) \) replaced by the identity function in \( u \). It holds with probability at least \( 1 - 2\delta \) that for any \( \gamma \in \Gamma_{01}, \)

\[
E_{P_n} h_{\gamma, \hat{\mu}}(x) - E_{P} h_{\gamma, \hat{\mu}}(x)
\leq (2p)^{-1/2} \left\{ C_{sg, 12} \sqrt{\frac{2qV_g}{n} + \frac{q \log(\delta^{-1})}{n}} \right\}
\leq (2p)^{-1/2} \left\{ C_{sg, 12} C_{\text{rad}3} \sqrt{\frac{16p}{n} + \frac{2p \log(\delta^{-1})}{n}} \right\},
\]  

(152)

as shown in Proposition 47. Combining the inequalities (150)–(152) leads to the desired result.

\section*{Proposition 56}

In the setting of Proposition 26, it holds with probability at least \( 1 - 2\delta \) that for any \( \gamma \in \Gamma_{20}, \)

\[
K_{HG}(P_n, P_{\theta'}; h_{\gamma, \hat{\mu}})
\geq \left\{ E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) - E_{P_{\theta'}} h_{\gamma, \hat{\mu}}(x) \right\} - \epsilon - \sqrt{n} \lambda_{32}(4q)^{-1/2}
\]

92
where $q = p(1 - p)$, and, with $C_{rad6} = C_{sg,12} C_{rad3}$,

$$
\lambda_{32} = C_{rad6} \sqrt{\frac{12(p - 1)}{n}} + \sqrt{\frac{(p - 1) \log(\delta^{-1})}{n}}.
$$

**Proof** For any $\gamma \in \Gamma_{20} \subset \Gamma_{rp2}$, because $pen_2(\gamma) = (4q)^{-1/2}$ and $\gamma_0 = 0$, we have $pen_2(\beta) = 2pen_2(\gamma) = q^{-1/2}$. Hence $h_{\gamma}(x) = h_{rp2,\beta}(x) \in [-1,1]$ by the boundedness of the ramp function in $[0,1]$ and the Cauchy–Schwartz inequality, $\|\beta_2\|_1 \leq q^{1/2}\|\beta_2\|_2$. Then the mean-centered version $h_{\gamma,\tilde{\mu}}(x)$, with $E_{P_\theta} h_{\gamma,\tilde{\mu}}(x) = 0$, is also bounded in $[-1,1]$. Moreover, such $h_{\gamma,\tilde{\mu}}(x)$ can be expressed in the form $pen_2(\beta) w^T \{g(x - \tilde{\mu}) - \eta_0\}$, 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 $ramp(x_i)ramp(x_j)$ for $1 \leq i \neq j \leq p$, and $\eta_0 = E_{P_\theta} g(x - \tilde{\mu}) \in [0,1]^q$.

Next, $K_{HG}(P_n, P_\theta; h_{\gamma,\tilde{\mu}})$ can be bounded as

$$
K_{HG}(P_n, P_\theta; h_{\gamma,\tilde{\mu}}) \geq K_{HG}(P_\epsilon, P_\theta; h_{\gamma,\tilde{\mu}}) - |K_{HG}(P_n, P_\theta; h_{\gamma,\tilde{\mu}}) - K_{HG}(P_\epsilon, P_\theta; h_{\gamma,\tilde{\mu}})|.
$$

For any $\gamma \in \Gamma_{20}$, because $h_{\gamma,\tilde{\mu}}(x) \in [-1,1]$, we have $\min(h_{\gamma,\tilde{\mu}}(x),1) = h_{\gamma,\tilde{\mu}}(x)$ and $\min(-h_{\gamma,\tilde{\mu}}(x),1) = -h_{\gamma,\tilde{\mu}}(x)$. Then the hinge term $K_{HG}(P_n, P_\theta; h_{\gamma,\tilde{\mu}})$ reduces to a moment matching term and can be lower bounded as follows:

$$
K_{HG}(P_\epsilon, P_\theta; h_{\gamma,\tilde{\mu}}) = E_{P_\epsilon} \min(h_{\gamma,\tilde{\mu}}(x),1) + E_{P_\theta} \min(-h_{\gamma,\tilde{\mu}}(x),1) = E_{P_\epsilon} h_{\gamma,\tilde{\mu}}(x) - E_{P_\theta} h_{\gamma,\tilde{\mu}}(x) = \epsilon E_{P_\epsilon} h_{\gamma,\tilde{\mu}}(x) + (1 - \epsilon) E_{P_\theta} h_{\gamma,\tilde{\mu}}(x) - E_{P_\theta} h_{\gamma,\tilde{\mu}}(x) \geq -\epsilon + \left\{E_{P_\epsilon} h_{\gamma,\tilde{\mu}}(x) - E_{P_\theta} h_{\gamma,\tilde{\mu}}(x)\right\}.
$$

Similarly, the absolute difference term in (153) can be simplified as follows:

$$
|K_{HG}(P_n, P_\theta; h_{\gamma,\tilde{\mu}}) - K_{HG}(P_\epsilon, P_\theta; h_{\gamma,\tilde{\mu}})| = \left|\left\{E_{P_n} h_{\gamma,\tilde{\mu}}(x) - E_{P_\theta} h_{\gamma,\tilde{\mu}}(x)\right\} - \left\{E_{P_\epsilon} h_{\gamma,\tilde{\mu}}(x) - E_{P_\theta} h_{\gamma,\tilde{\mu}}(x)\right\}\right|
$$

We apply Lemma 45 with $b = 1$, $g(x) \in [0,1]^q$ and $\eta_0$ defined above, and $f'(e^u)$ and $f'(e^u)$ replaced by the identity function in $u$. It holds with probability at least $1 - 2\delta$ that for any $\gamma \in \Gamma_{20}$,

$$
|E_{P_n} h_{\gamma,\tilde{\mu}}(x) - E_{P_\epsilon} h_{\gamma,\tilde{\mu}}(x)| \leq (4q)^{-1/2} \left\{C_{sg,12} \sqrt{\frac{2q V_{\theta}}{n}} + \sqrt{\frac{q \log(\delta^{-1})}{n}}\right\}
$$

$$
= \sqrt{p}(4q)^{-1/2} \left\{C_{sg,12} C_{rad3} \sqrt{\frac{12(p - 1)}{n}} + \sqrt{\frac{(p - 1) \log(\delta^{-1})}{n}}\right\},
$$

as shown in Proposition 49. Combining the inequalities (153)–(155) leads to the desired result. \[\square\]
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

$$K_f(P_\epsilon, P_{\hat{\theta}}; h) \leq -f'(1 - \epsilon_0)\epsilon.$$

(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

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

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

$$K_f(P_\epsilon, P_{\hat{\theta}}; h) = \mathbb{E}_{P_\epsilon} f'(e^{h(x)}) - \mathbb{E}_{P_{\hat{\theta}}} f^\#(e^{h(x)})$$

$$= \mathbb{E}_{P_\epsilon} T(h(x)) - \mathbb{E}_{P_{\hat{\theta}}} f^\#\{T(h(x))\},$$

where $T(u) = f'(e^u)$ 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, $\mathbb{E}_{P_\epsilon} f'(e^{h_{\gamma, \mu}(x)}) - \mathbb{E}_{P_{\hat{\theta}}} G(h_{\gamma, \mu}(x))$. By the definition of a solution to alternating optimization (see Remark 1), we have

$$L_G((\hat{\mu}, \hat{\Sigma}), \hat{\gamma}) \leq L_G((\mu^*, \Sigma^*), \gamma),$$

$$L_G((\hat{\mu}, \hat{\Sigma}), \gamma) \leq L_G((\hat{\mu}, \Sigma^*), \hat{\gamma}),$$

that is, the generator loss at $\hat{\theta}$ is less than that at any $\theta$, with the discriminator parameter fixed at $\hat{\gamma}$. The preceding inequalities can be written out as

$$\mathbb{E}_{P_\epsilon} T(e^{h_{\gamma, \hat{\mu}}(x)}) - \mathbb{E}_{P_{\hat{\theta}}} G(h_{\gamma, \hat{\mu}}(x)) \leq \mathbb{E}_{P_\epsilon} T(e^{h_{\gamma, \mu^*}(x)}) - \mathbb{E}_{P_{\mu^*, \Sigma}} G(h_{\gamma, \mu^*}(x)),$$

and

$$\mathbb{E}_{P_\epsilon} T(e^{h_{\gamma, \hat{\mu}}(x)}) - \mathbb{E}_{P_{\hat{\theta}}} G(h_{\gamma, \hat{\mu}}(x)) \leq \mathbb{E}_{P_{\hat{\mu}, \Sigma}} T(e^{h_{\gamma, \hat{\mu}}(x)}) - \mathbb{E}_{P_{\hat{\mu}, \Sigma}} G(h_{\gamma, \hat{\mu}}(x)).$$

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 $\mathbb{E}_{P_{\hat{\theta}}} G(h_{\gamma, \hat{\mu}}(x)) = \mathbb{E}_{P_{\mu^*, \Sigma}} G(h_{\gamma, \mu^*}(x))$ because both are equal to $\mathbb{E}_{P_{\hat{\theta}}} G(h_{\gamma, 0}(x))$. In the second inequality, the term $\mathbb{E}_{P_\epsilon} T(e^{h_{\gamma, \hat{\mu}}(x)})$ is on both sides. Then the two inequalities yield

$$\mathbb{E}_{P_\epsilon} T(e^{h_{\gamma, \hat{\mu}}(x)}) \leq \mathbb{E}_{P_{\hat{\theta}}} T(e^{h_{\gamma, \mu^*}(x)}),$$

$$-\mathbb{E}_{P_{\hat{\theta}}} G(h_{\gamma, \hat{\mu}}(x)) \leq -\mathbb{E}_{P_{\mu^*, \Sigma}} G(h_{\gamma, \mu^*}(x)).$$
Now we are ready to derive an upper bound for $K_f(P_\epsilon, P_{\hat{\theta}}; h_{\hat{\gamma}, \hat{\mu}})$:

$$
K_f(P_\epsilon, P_{\hat{\theta}}; h_{\hat{\gamma}, \hat{\mu}}) = E_{\theta} T(h_{\hat{\gamma}, \hat{\mu}}) - E_{P_{\hat{\theta}}} f^*(T(G^{-1}(G(h_{\hat{\gamma}, \hat{\mu}^*})))) \\
\leq E_{\theta} T(h_{\hat{\gamma}, \hat{\mu}}) - f^*(T(G^{-1}(E_{P_{\hat{\theta}}} G(h_{\hat{\gamma}, \hat{\mu}})))) \\
\leq (1-\epsilon)E_{P_{\hat{\theta}}}(T(h_{\hat{\gamma}, \hat{\mu}}) - f^*(T(G^{-1}(E_{P_{\hat{\theta}}} G(h_{\hat{\gamma}, \hat{\mu}}))))) \\
\leq (1-\epsilon)f(E_{P_{\hat{\theta}}}(h_{\hat{\gamma}, \hat{\mu}}) - f^*(T(E_{P_{\hat{\theta}}} G(h_{\hat{\gamma}, \hat{\mu}})))) \\
\leq f(1-\epsilon) \leq -f'(1-\epsilon_0)\epsilon.
$$

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 $E_{P_{\hat{\theta}}} G(h_{\hat{\gamma}, \hat{\mu}}) = E_{P_{\hat{\theta}}} G(h_{\hat{\gamma}, \hat{\mu}^*}(x))$ and drop the $E_Q$ term because $T \leq 0$. 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

$$(1-\epsilon)s - f^*(s) \leq f(1-\epsilon) \leq -f'(1-\epsilon_0)\epsilon,$$

with $s$ set to $E_{P_{\hat{\theta}}}(T(h_{\hat{\gamma}, \hat{\mu}^*})$.

(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):

$$
K_f(P_n, P_{\hat{\theta}}; h_{\hat{\gamma}}) = \epsilon E_{\theta} T(h_{\hat{\gamma}, \hat{\mu}}) + (1-\epsilon)E_{P_{\hat{\theta}^*}} T(h_{\hat{\gamma}, \hat{\mu}}) - E_{P_{\hat{\theta}}} f^*(T(h_{\hat{\gamma}, \hat{\mu}})) \\
\leq (1-\epsilon)f(E_{P_{\hat{\theta}^*}}(h_{\hat{\gamma}, \hat{\mu}^*}) - f^*(T(E_{P_{\hat{\theta}^*}} G(h_{\hat{\gamma}, \hat{\mu}^*})))) + R_1|E_{P_{\hat{\theta}^*}}(h_{\hat{\gamma}, \hat{\mu}^*} - E_{P_{\hat{\theta}}}(h_{\hat{\gamma}, \hat{\mu}^*})| \\
\leq f(1-\epsilon) + R_1|E_{P_{\hat{\theta}^*}}(h_{\hat{\gamma}, \hat{\mu}^*} - E_{P_{\hat{\theta}}}(h_{\hat{\gamma}, \hat{\mu}^*})| \\
\leq f(1-\epsilon) + R_1|E_{P_{\hat{\theta}^*}}(h_{\hat{\gamma}, \hat{\mu}^*} - E_{P_{\hat{\theta}}}(h_{\hat{\gamma}, \hat{\mu}^*})|.
$$

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 $h : \mathbb{R}^p \rightarrow \mathbb{R}$, we have

$$
K_{HG}(P_\epsilon, P_{\hat{\theta}}; h) \leq 2\epsilon.
$$

(ii) If $\epsilon = n^{-1} \sum_{i=1}^n U_i \in [0,1]$, then for any function $h : \mathbb{R}^p \rightarrow \mathbb{R}$, we have

$$
K_{HG}(P_n, P_{\hat{\theta}}; h) \leq 2\epsilon + |E_{P_{\hat{\theta}^*}} h(x) - E_{P_{\hat{\theta}}}(h(x))|, \tag{162}
$$

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

95
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:

$$K_{HG}(P_\epsilon, P_{\hat{\theta}}; h_{\hat{\gamma}, \hat{\mu}})$$

$$\leq \epsilon + (1 - \epsilon) \min(E_{P_\theta^*} h_{\hat{\gamma}, \hat{\mu}^*}, 1) + \min(-E_{P_\theta^*} h_{\hat{\gamma}, \hat{\mu}^*}, 1)$$

(163)

$$\leq 2\epsilon.$$  

(164)

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 $\epsilon$ term is a result of the fact that $\min(u,1) \leq 1$. Inequality (164) is by the same argument used in Lemma 51:

$$\epsilon \leq (1 - \epsilon) \{\min(u,1) + \min(-u,1)\}$$

$$\leq \epsilon,$$

with $u$ set to be $h_{\hat{\gamma}, \hat{\mu}^*}$.

(ii) To derive an upper bound for $K_{HG}(P_n, P_{\hat{\theta}}; h_{\hat{\gamma}})$, we first argue similarly as in part (i):

$$K_{HG}(P_n, P_{\hat{\theta}}; h_{\hat{\gamma}, \hat{\mu}})$$

$$= \hat{\epsilon} E_Q \min(h_{\hat{\gamma}, \hat{\mu}}, 1) + (1 - \hat{\epsilon}) E_{P_\theta^*, n} \min(h_{\hat{\gamma}, \hat{\mu}^*}, 1) + E_{P_\theta} \min(-h_{\hat{\gamma}, \hat{\mu}}, 1)$$

$$\leq \hat{\epsilon} + (1 - \hat{\epsilon}) \min(E_{P_\theta^*, n} h_{\hat{\gamma}, \hat{\mu}^*}, 1) + \min(-E_{P_\theta^*, \Sigma^*} h_{\hat{\gamma}, \hat{\mu}^*}, 1).$$

Then we use the 1-Lipschitz property of $\min(-u,1)$ and obtain

$$\epsilon \leq (1 - \epsilon) \min(E_{P_\theta^*, n} h_{\hat{\gamma}, \hat{\mu}^*}, 1) + \min(-E_{P_\theta^*, \Sigma^*} h_{\hat{\gamma}, \hat{\mu}^*}, 1)$$

$$\leq (1 - \epsilon) \min(E_{P_\theta^*, n} h_{\hat{\gamma}, \hat{\mu}^*}, 1) + \min(-E_{P_\theta^*, n} h_{\hat{\gamma}, \hat{\mu}^*}, 1) + |E_{P_\theta^*, n} h_{\hat{\gamma}, \hat{\mu}^*} - E_{P_\theta} h_{\hat{\gamma}, \hat{\mu}^*}|$$

$$\leq \hat{\epsilon} + |E_{P_\theta^*, n} h_{\hat{\gamma}, \hat{\mu}^*} - E_{P_\theta} h_{\hat{\gamma}, \hat{\mu}^*}|.$$  

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 $f''(t) = \frac{2}{(1+t)^2} g_0(\frac{2t}{1+t})$, 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 $L_g(P_*, P_0; q_\gamma) = K_f(P_*, P_0; h_\gamma)$. Denote $e^{h_\gamma(x)}$ by $t$ and $q_\gamma(x)$ by $q$.

Then we have

\begin{align*}
    K_f(P_*, P_0; h_\gamma) &= E_{P_0} f''(t) - E_{P_0} \left\{ tf'(t) - f(t) \right\} \\
    &= E_{P_0} \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\} - E_{P_0} \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\} \\
    &= E_{P_0} \left\{ (1-q) g_0' (2q) + \frac{1}{2} g_0 (2q) \right\} - E_{P_0} \left\{ q g_0' (2q) - \frac{1}{2} g_0 (2q) \right\} \\
    &= E_{P_0} \left\{ \frac{1}{2} g' (q) + \frac{1}{2} g (q) - \frac{1}{2} g \left( \frac{1}{2} \right) \right\} - E_{P_0} \left\{ q g' (q) - \frac{1}{2} g (q) + \frac{1}{2} g \left( \frac{1}{2} \right) \right\} \\
    &= \frac{1}{2} \left\{ E_{P_0} S_g (q, 1) - E_{P_0} S_g (q, 0) \right\} - g \left( \frac{1}{2} \right). \\
\end{align*}

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

\[ \int qf \left( \frac{p}{q} \right) = \int \frac{p+q}{2} g_0 \left( \frac{2p}{p+q} \right), \]

which implies that $D_{g_0}(P_\ast\| (P_\ast + P_0)/2) = D_f(P_\ast\| P_0)$.

### 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 $X \sim N(0, \sigma^2)$ and $\xi \in \mathbb{R}$, we have

\[ \mathbb{E}(X - \xi)_+ \leq \frac{\sigma}{\sqrt{2\pi}} + |\xi|, \]

\[ \mathbb{E}[(X - \xi)_+]^2 \leq \sigma^2 + \xi^2. \]

**Proof** The second result is immediate: $\mathbb{E}[(X - \xi)_+]^2 \leq \mathbb{E}(X - \xi)^2 = \sigma^2 + \xi^2$. The first result can be shown as follows:

\[ \mathbb{E}(X - \xi)_+ = \int_{\xi}^{\infty} (x - \xi) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \]

\[ = \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2\sigma^2}} - \xi \int_{\xi}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \]

\[ \leq \frac{\sigma}{\sqrt{2\pi}} + |\xi|. \]

The last inequality holds because $0 \leq \int_{\xi}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \leq 1$. \hfill \blacksquare
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

$$\Delta_n^C(x_1, \ldots, x_n) = \#\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\},$$

that is, $\Delta_n^C(x_1, \ldots, x_n)$ is the number of subsets of $\{x_1, \ldots, x_n\}$ picked out by the collection $\mathcal{C}$. We say that a subset $\{x_i, \ldots, x_j\} \subseteq \{x_1, \ldots, x_n\}$ 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 $\mathcal{C}$ if $\{x_i, \ldots, x_j\} = C \cap \{x_1, \ldots, x_n\}$. Moreover, define

$$m^C(n) = \max_{x_1, \ldots, x_n} \Delta_n^C(x_1, \ldots, x_n),$$

and the Vapnik–Chervonenkis (VC) index of $\mathcal{C}$ as (van der Vaart and Wellner, 1996)

$$V(\mathcal{C}) = \inf\{n \geq 1 : m^C(n) < 2^n\},$$

where the infimum over the empty set is taken to be infinity.

The subgraph of a function $f : \mathcal{X} \rightarrow \mathbb{R}$ 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

$$f_b(x_2) - f_b(x_1) \geq (-1)(x_2 - x_1) \geq t_2 - t_1.$$

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

$$f_b(x_1) - f_b(x_2) \geq (-1)(x_1 - x_2) > -(t_2 - t_1).$$
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 \( G_f \).

Combining the preceding cases shows that \( m^{G_F}(2) < 2^2 \) and \( V(F) = 2 \). Moreover, the class of functions \( \{f_b(-x) : b \in \mathbb{R}\} \), denoted as \( \overline{F} \), admits a one-to-one correspondence with \( F \). A subset of \( \{(x_1, t_1), (x_2, t_2)\} \) is picked up by \( G_f \) if and only if the corresponding subset of \( \{(-x_1, t_1), (-x_2, t_2)\} \) is picked up by \( G_{\overline{F}} \). Hence \( V(\overline{F}) = V(F) = 2 \). \[\]

**Lemma 61** For \( F = \{f(x) \equiv b : b \in \mathbb{R}\} \), we have that \( V(F) = 2 \). 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 \( t_1 \leq t_2 \). Then the singleton \( \{(x_2, t_2)\} \) can never be picked up by \( G_f \) for any \( f \in F \), and hence \( m^{G_F}(2) < 2^2 \) and \( V(F) = 2 \). In fact, if \( (x_2, t_2) \) is in the subgraph of \( f(x) \equiv b \), then \( t_2 < b \). As a result, \( t_1 \leq t_2 < b \), indicating that \( (x_1, t_1) \) is also in the subgraph. \[\]

### D.3 Lipschitz functions of Gaussian vectors

Say that a function \( g : \mathbb{R}^p \rightarrow \mathbb{R}^m \) is \( L \)-Lipschitz if \( \|g(x_1) - g(x_2)\|_2 \leq L \|x_1 - x_2\|_2 \) for any \( x_1, x_2 \in \mathbb{R}^p \).

**Lemma 62** Let \( X \sim N_p(\mu, \Sigma) \), and \( g : \mathbb{R}^p \rightarrow \mathbb{R}^m \) be an \( L \)-Lipschitz function.

(i) For any vector \( w \in \mathbb{R}^m \) with \( \|w\|_2 = 1 \), we have

\[
\text{E} \left[ \{w^T(g(X) - Eg(X))\}^2 \right] \leq 2C_{sg,12}^2 L^2 \|\Sigma\|_{op},
\]

where \( C_{sg,12} \) is the universal constant from Lemma 67. Hence we have

\[
\|\text{Var} g(X)\|_{op} \leq 2C_{sg,12}^2 L^2 \|\Sigma\|_{op}.
\]

(ii) For any symmetric matrix \( A \in \mathbb{R}^{m \times m} \) with \( \|A\|_F = 1 \), we have

\[
\text{E} \left[ \{g(X) - Eg(X)\}^T A (g(X) - Eg(X)) \right] \leq 4C_{sg,12}^4 m L^4 \|\Sigma\|_{op}.
\]

**Proof** (i) By Boucheron et al. (2013), Theorem 5.6, it can be shown that for any \( L \) unit vector \( w \), \( w^T(g(X) - Eg(X)) \) is sub-gaussian with tail parameter \( L \|\Sigma\|_{op}^{1/2} \). See the proof of Lemma 43(i) for a similar argument. Then by Lemma 67, \( E[\{w^T(g(X) - Eg(X))\}^2] \leq 2C_{sg,12}^2 L^2 \|\Sigma\|_{op} \).

(ii) Consider an eigen-decomposition \( A = \sum_{j=1}^m \lambda_j w_j w_j^T \), where \( \lambda_j \)'s are eigenvalues and \( w_j \)'s are the eigenvectors with \( \|w_j\|_2 = 1 \). Denote \( g = g(X) \) and \( \tilde{g} = g - Eg \). Then

\[
\tilde{g}^T A \tilde{g} = \sum_{j=1}^m \lambda_j (w_j^T \tilde{g})^2 \]

and

\[
(\tilde{g}^T A \tilde{g})^2 \leq \left( \sum_{j=1}^m \lambda_j^2 \right) \left( \sum_{j=1}^m (w_j^T \tilde{g})^4 \right) \leq 4C_{sg,12}^4 m L^4 \|\Sigma\|_{op}^2.
\]
The first step uses the Cauchy–Schwartz inequality, and the second step uses the fact that $\sum_{j=1}^{m} \lambda_j^2 = \|A\|_F^2 = 1$ and $E[(w_j^T \tilde{g})^4] \leq 4C_{sg,12}^2 L^4\|\Sigma\|_{op}^2$ by Lemma 67 because $w_j^T \tilde{g}_j$ is sub-gaussian with tail parameter $L\|\Sigma\|_{op}^{1/2}$ for each $j$. \hfill \square

The following result provides a 4th-moment bound which depends linearly on $L^2\|\Sigma\|_{op}$, under a boundedness condition in addition to the Lipschitz condition.

**Lemma 63** Let $X \sim N_p(\mu, \Sigma)$, 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 $\|A\|_F = 1$, we have

$$E \left[ \{(g(X) - Eg(X))^T A(g(X) - Eg(X))\}^2 \right] \leq 2C_{sg,12}^2 m L^2 \|\Sigma\|_{op}.$$  

(ii) For any matrix $A \in \mathbb{R}^{m \times m}$ with $\|A\|_F = 1$, we have

$$E \left[ \{g^T(X)Ag(X) - Eg^T(X)Ag(X)\}^2 \right] \leq 20C_{sg,12}^2 m L^2 \|\Sigma\|_{op}.$$

**Proof** (i) Denote $g = g(X)$ and $\tilde{g} = g - Eg$. Then each component of $\tilde{g}$ is contained in $[-1,1]$ by the boundedness of $g$. The variable $(\tilde{g}^T A \tilde{g})^2$ can be bounded as follows:

$$(\tilde{g}^T A \tilde{g})^2 = \text{tr}(\tilde{g}^T A \tilde{g})^2 = \text{tr}(A \tilde{g}^T A \tilde{g})^2 \leq \text{tr}(A \tilde{g}^T A)^2 \|\tilde{g}\|_2^2 \|\tilde{g}\|_2 = m \text{tr}(\tilde{g}^T A)^2 \|\tilde{g}\|_2^2.$$  

Line (168) follows from von Neumann’s trace equality. Line (169) uses the fact that $\|\tilde{g}\|_2 \leq m$, because $w^T \tilde{g} u = (w^T \tilde{g})^2 \leq \|w\|_2^2 \|\tilde{g}\|_2^2 \leq m \|w\|_2^2$ for any $w \in \mathbb{R}^m$, by the boundedness of $\tilde{g}$. Then the desired result follows because

$$E \text{tr}(A^T A \tilde{g} \tilde{g}^T) = \text{tr}(A^T A \text{Var}(g)) \leq \text{tr}(A^T A \|\text{Var}(g)\|_{op}) \leq 2C_{sg,12}^2 m L^2 \|\Sigma\|_{op}. \quad (170)$$

Line (170) also follows from von Neumann’s trace equality. Line (171) follows because $\text{tr}(A^T A) = \|A\|_F^2 = 1$ and $\|\text{Var}(g)\|_{op} \leq 2C_{sg,12}^2 m L^2 \|\Sigma\|_{op}$ by Lemma 62(i), with $g$ being an $L$-Lipschitz function.

(ii) The difference $g^T Ag - Eg^T Ag$ can be expressed in terms of the centered variables as $g^T Ag - Eg^T Ag = (\tilde{g}^T A \tilde{g} - Eg^T A \tilde{g}) + 2(Eg)^T A \tilde{g}$.

Then

$$(g^T Ag - Eg^T Ag)^2 \leq 2(\tilde{g}^T A \tilde{g} - Eg^T A \tilde{g})^2 + 8((Eg)^T A \tilde{g})^2 \leq 2(\tilde{g}^T A \tilde{g} - Eg^T A \tilde{g})^2 + 8\|Eg\|_2^2 \|A \tilde{g}\|_2^2.$$  

The expectation of the first term on (172) can be bounded using (i) as

$$2E\{(\tilde{g}^T A \tilde{g} - Eg^T A \tilde{g})^2\} = 2E\{(\tilde{g}^T A \tilde{g})^2\} - 2(Eg^T A \tilde{g})^2 \leq 2E\{(\tilde{g}^T A \tilde{g})^2\} \leq 4mC_{sg,12}^2 m L^2 \|\Sigma\|_{op}.$$
The expectation of the second term on (172) can be bounded as
\[ 8\|Eg\|_2^2 \|A\|_2^2 = 8\|Eg\|_2^2 \text{Etr}(A^T A \tilde{g} \tilde{g}^T) \]
\[ \leq 16mC^2_{\text{sg},12} L^2 \|\Sigma\|_{\text{op}}. \]
by inequality (171) and the fact that \( \|Eg\|_2^2 \leq m \). Combining the preceding two bounds yields the desired result.

\[ \square \]

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 \( X_1 = \mu_1 + D_1 Z \) and \( X_2 = \mu_2 + D_2 Z \), 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 \( D_1 = \text{diag}(d_1) \) and \( D_2 = \text{diag}(d_2) \) with \( d_1, d_2 \in \mathbb{R}^p \). Then for any matrix \( A \in \mathbb{R}^{m \times m} \) with \( \|A\|_F = 1 \),
\[ |Eg^T(X_1)Ag(X_1) - Eg^T(X_2)Ag(X_2)| \]
\[ \leq 2\sqrt{m}L^2 \Delta^2 + 2\sqrt{2}L \Delta (E\|g\|_2^2)^{1/2}, \]
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 \( g_1^T A g_1 - g_2^T A g_2 \) can be decomposed as
\[ g_1^T A g_1 - g_2^T A g_2 \]
\[ = (g_1 - g_2)^T A (g_1 - g_2) + 2(g_1 - g_2)^T A g_2. \quad (173) \]
The expectation of the first term on (173) can be bounded as
\[ |E(g_1 - g_2)^T A (g_1 - g_2)| = |\text{tr}(AV)| \]
\[ \leq \sum_{j=1}^{m} s_j(A) \|V\|_{\text{op}} \quad (174) \]
\[ \leq 2\sqrt{m}L^2 \left( \|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2 \right), \quad (175) \]
where \( s_1(A), \ldots, s_m(A) \) are the singular values of \( A \), and \( V = E\{(g_1 - g_2)(g_1 - g_2)^T\} \). Line (174) follows from von Neumann’s trace inequality. Line (175) follows because \( \sum_{j=1}^{m} s_j(A) \leq \sqrt{m} \left( \sum_{j=1}^{m} s_j^2(A) \right)^{1/2} = \sqrt{m} \|A\|_F = \sqrt{m} \) with \( \|A\|_F = 1 \) and \( \|V\|_{\text{op}} \leq 2L^2(\|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2) \), which can be shown as follows. For any \( L_2 \) unit vector \( w \), we have
\[ w^T V w = E \left[ (w^T (g_1 - g_2))^2 \right] \leq E\|g_1 - g_2\|_2^2 \]
\[ \leq L^2 \|\mu_1 + D_1 Z - (\mu_2 + D_2 Z)\|_2^2 \]
\[ \leq 2L^2 \left( \|\mu_1 - \mu_2\|_2^2 + E\|(D_1 - D_2) Z\|_2^2 \right) \]
\[ \leq 2L^2 \left( \|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2 \right). \quad (176) \]
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

$$|E((g_1 - g_2)^T A g_2| \leq E \|g_1 - g_2\|_2 \|A g_2\|_2$$

$$\leq E^{1/2} (\|g_1 - g_2\|_2^3) E^{1/2} (\|A g_2\|_2^3)$$

$$\leq \sqrt{2} L (\|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2)^{1/2} (E \|g_2\|_2^3)^{1/2}.$$  \hspace{1cm} (177)

Line (177) uses the fact that $E\{\|g_1 - g_2\|_2^3\} \leq 2L^2 (\|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2)$ based on (176) and the following argument:

$$E \|A g_2\|_2^2 \leq E (\|A\|_{op}^2 \|g_2\|_2^2) \leq E \|g_2\|_2^2,$$

where the last step follows because $\|A\|_{op} \leq \|A\|_F = 1$. Combining (173), (175), and (177) yields the desired result. \hfill \blacksquare

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 $\|A\|_F = 1$,

$$|E g^T (X_1) A g(X_1) - E g^T (X_2) A g(X_2)| \leq 2 \sqrt{2} m L \Delta,$$

where $\Delta^2 = \|\mu_1 - \mu_2\|_2^2 + \|d_1 - d_2\|_2^2$.

**Proof** The difference $g_1^T A g_1 - g_2^T A g_2$ can also be decomposed as

$$g_1^T A g_1 - g_2^T A g_2 = (g_1 - g_2)^T A g_1 + (g_1 - g_2)^T A g_2.$$

By (177), both of the two terms on the right-hand side can be bounded in absolute values by $\sqrt{2} L \Delta (E \|g_2\|_2^3)^{1/2}$ and hence by $\sqrt{2} m L \Delta$, because $E \|g_2\|_2^3 \leq m$ by the componentwise boundedness of $g_1$ and $g_2$. \hfill \blacksquare

**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 $\alpha_1 \geq \cdots \geq \alpha_m \geq 0$ and $\beta_1 \geq \cdots \geq \beta_m \geq 0$ respectively,

$$|\text{tr}(AB)| \leq \sum_{j=1}^{m} \alpha_j \beta_j.$$

As a direct consequence, if $A$ is symmetric and non-negative definite, then

$$|\text{tr}(AB)| \leq \text{tr}(A) \|B\|_{op}$$

This follows because the singular values $\alpha_i$’s are also the eigenvalues of $A$ and hence $\text{tr}(A) = \sum_{j=1}^{m} \alpha_j$ for a symmetric and nonnegative definite matrix $A$, and $\|B\|_{op} = \max_{i=j, \ldots, m} \beta_j$ by the definition of $\|B\|_{op}$.  

102
E.2 Sub-gaussian 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 $C_{sg, ij} > 0$ such that the $C_{sg, ij}^{-1} K_j \leq K_i \leq C_{sg, ij} K_j$ for all $1 \leq i \neq j \leq 4$, where $K_i$ is the parameter appearing in property (i).

(i) $P(|Y| > t) \leq 2 \exp\left(-\frac{t^2}{2K_1^2}\right)$ for any $t > 0$.

(ii) $E^{1/p}(|Y|^p) \leq K_2 \sqrt{p}$ for any $p \geq 1$.

(iii) $E \exp(Y^2/K_1^2) \leq 2$.

If $EY = 0$, then properties (1)–(3) are also equivalent to the following one.

(iv) $E \exp(sY) \leq \exp\left(\frac{K_2^2 s^2}{2}\right)$, 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

$$P(|Y| > b + t) \leq 2e^{-\frac{t^2}{2K^2}} \quad \text{for any } t > 0.$$ 

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

$$P(|Y| > y) \leq 2 \exp\left\{-\frac{(y - b)^2}{2K^2}\right\} \leq 2 \exp\left\{-\frac{y^2}{2(K + b)^2}\right\}.$$ 

Second, if $y \leq K + b$, then

$$P(|Y| > y) \leq 1 \leq 2e^{-\frac{t^2}{2K^2}} \leq 2 \exp\left\{-\frac{y^2}{2(K + b)^2}\right\}.$$ 

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

$$P\left(\left|\frac{1}{n} \sum_{i=1}^{n} Y_i\right| > C_{sg5} t\right) \leq 2 \exp\left(-\frac{nt^2}{2K^2}\right) \quad \text{for any } t > 0.$$ 

where $C_{sg5} = C_{sg, 14}$ 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

\[
E \max_{i=1,\ldots,n} |Y_i| \leq C_{sg6}K\sqrt{\log(2n)},
\]

where \(C_{sg6} > 0\) is a universal constant.

Say that \(Y \in \mathbb{R}^d\) is a sub-gaussian random vector with tail parameter \(K\) if \(w^T Y\) 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 \(EY = 0\) is a sub-gaussian random vector with tail parameter \(K\). Then for any \(t > 0\), we have that with probability at least \(1 - e^{-t}\),

\[
\|Y\|_2 \leq C_{sg7}K(\sqrt{d} + \sqrt{t}),
\]

where \(C_{sg7} > 0\) 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 \(1 - 2e^{-t}\),

\[
\left\| \frac{1}{n} \sum_{i=1}^n Y_i Y_i^T - \Sigma \right\|_{op} \leq C_{sg8}K^2 \left(\sqrt{\frac{d + t}{n}} + \frac{d + t}{n}\right),
\]

where \(\Sigma = E(YY^T)\) 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 \(C_{sx,ij} > 0\) such that the \(C_{sx,ij}^{-1} K_j \leq K_i \leq C_{sx,ij} K_j\) for all \(1 \leq i \neq j \leq 4\), where \(K_i\) is the parameter appearing in property (i).

(i) \(P(|Y| > t) \leq 2 \exp(-\frac{t^2}{K_1})\) for any \(t > 0\).

(ii) \(E^{1/p}(|Y|^p) \leq K_2 p\) for any \(p \geq 1\).

(iii) \(E \exp(|Y|/K_3) \leq 2\).

If \(EY = 0\), then properties (i)–(iii) are also equivalent to the following one.

(iv) \(E \exp(sY) \leq \exp\left(\frac{K_4 s^2}{2}\right),\) 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 \( C_{sx5} K_1 K_2 \), where \( C_{sx5} > 0 \) 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 \( Y - \mathbb{E} Y \) is sub-exponential random variable with tail parameter \( C_{sx6} K \), where \( C_{sx6} > 0 \) 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

\[
P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Y_i \right| > C_{sx7} K \left( \sqrt{\frac{t}{n}} \vee \frac{t}{n} \right) \right\} \leq 2e^{-t} \quad \text{for any } t > 0.
\]

where \( C_{sx7} > 0 \) 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 \( Ef(X_1) < \infty \) for all \( f \in \mathcal{F} \). Then we have

\[
E \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - Ef(X_i) \right\} \leq 2E \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right\},
\]

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

\[
E \sup_{f \in \mathcal{F}} \left\{ Ef(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\}.
\]
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

\[
E \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - Ef(X) \right\} = E_{X_i} \left[ \sup_{f \in \mathcal{F}} E_{X'_i} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X'_i) \right\} \right] 
\]

(178)

\[
\leq E_{X_i, X'_i, \epsilon_i} \left[ \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(X'_i) \right\} \right] 
\]

(179)

\[
\leq E_{X_i, \epsilon_i} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right] + E_{X'_i, \epsilon_i} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} -\epsilon_i f(X'_i) \right] 
\]

\[
= 2E_{X_i, \epsilon_i} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right] .
\]

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 \(E \sup_{f \in \mathcal{F}} \left\{ Ef(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\} \).

\[\blacksquare\]

**Lemma 78** Let \(\phi\) be a function with a Lipschitz constant \(R\). Then in the setting of Lemma 77, we have

\[
E \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(f(X_i)) \right\} \leq E \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i Rf(X_i) \right\} .
\]

**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 \(\delta\) 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 \( \sup_{f \in F} \|f\|_n \leq 1 \) and \( \int_0^1 H^{1/2}(u, F, \| \cdot \|_n) \, du \leq \Psi_n(F) \), 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 \( t > 0 \),
\[
\Pr\left\{ Z_n(F)/C_{rad} > n^{-1/2}(\Psi_n(F) + t) \right\} \leq 2e^{-\frac{t^2}{2}},
\]
(180)
where \( C_{rad} > 0 \) 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 \( r \geq 1 \) and probability measure \( Q \), the \( L_r(Q) \) norm is defined as \( \|f\|_{r,Q} = (\int |f|^r \, dQ)^{1/r} \).

**Lemma 80** Let \( F \) be a VC class of functions such that \( \sup_{f \in F} |f| \leq 1 \). Then for any \( r \geq 1 \) and probability measure \( Q \), we have
\[
\mathcal{N}(u, F, \| \cdot \|_{r,Q}) \leq C_{vc} V(F) (16e)^{V(F)} u^{-r(V(F) - 1)}
\]
for any \( u \in (0, 1) \),
where \( V(F) \) denotes the VC index of \( F \) and \( C_{vc} \geq 1 \) 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 \( Z_n(F) \) is sub-gaussian with tail parameter \( C_{rad} n^{-1/2}(\Psi_n(F) + 1) \).

**Proof** By (180), the results follows from an application of Lemma 68. \( \blacksquare \)

**Corollary 82** In the setting of Lemma 79, the following results hold.

(i) If \( \sup_{f \in F} |f| \leq 1 \), then \( Z_n(F) \) is sub-gaussian with tail parameter \( C_{rad2} \sqrt{V(F)/n} \), where \( C_{rad2} = C_{rad}(1 + \sqrt{2 + \log(16C_{vc})} + \int_0^1 \sqrt{2 \log(u^{-1})} \, du) \).

(ii) Consider another two classes \( G \) and \( H \) of functions from \( X \) to \( \mathbb{R} \) in addition to \( F \), and let \( F_{com} = \{ f + g : f \in F, g \in G, h \in H \} \). If \( \sup_{f \in F_{com} \cup G \cup H} |f| \leq 1 \), then \( Z_n(F_{com}) \) is sub-gaussian with tail parameter \( C_{rad3} \sqrt{V(F) + V(G) + V(H)}/n \), where \( C_{rad3} = C_{rad}(1 + \sqrt{2 + \log(16C_{vc})} + \int_0^1 \sqrt{2 \log(3u^{-1})} \, du) \).

**Proof** (i) Take \( r = 2 \) and \( Q \) to be the empirical distribution on \( \{x_1, \ldots, x_n\} \). By Lemma 80, the entropy integral \( \int_0^1 H^{1/2}(u, F, \| \cdot \|_n) \, du \) can be upper bounded by
\[
\int_0^1 \log^{1/2}\left\{ V_{vc} V(F) (16e)^{V(F)} u^{-2(V(F) - 1)} \right\} \, du
= \int_0^1 \{ \log(C_{vc}) + \log V(F) + V(F) \log(16e) + 2(V(F) - 1) \log(u^{-1}) \}^{1/2} \, du
\leq \sqrt{V(F)} \int_0^1 \left\{ \sqrt{2 + \log(16C_{vc})} + \sqrt{2 \log(u^{-1})} \right\} \, du,
\]

107
using \( \log V(\mathcal{F}) \leq V(\mathcal{F}) \) for \( V(\mathcal{F}) \geq 1 \) and \( \sqrt{u_1 + u_2} \leq \sqrt{u_1} + \sqrt{u_2} \). 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 \( \mathcal{N}(u, \mathcal{F}_{\text{com}}; \| \cdot \|_n) \) 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

\[
\| \hat{f} - f \|_n \leq u/3, \quad \| \hat{g} - g \|_n \leq u/3, \quad \| \hat{h} - h \|_n \leq u/3.
\]

By the triangle inequality and \( \sup_{f \in \mathcal{F}, g \in \mathcal{G}} |f| \leq 1 \), we have

\[
\| \hat{f} \hat{g} + \hat{h} - f g - h \|_n
\leq \| (\hat{f} - f) \hat{g} \|_n + \| f (\hat{g} - g) \|_n + \| \hat{h} - h \|_n
\leq \| \hat{f} - f \|_n + \| \hat{g} - g \|_n + \| \hat{h} - h \|_n \leq u.
\]

This shows that \( \hat{\mathcal{F}}_{\text{com}} = \{ \hat{f} \hat{g} + \hat{h} : \hat{f} \in \hat{\mathcal{F}}, \hat{g} \in \hat{\mathcal{G}}, \hat{h} \in \hat{\mathcal{H}} \} \) is a \( u \)-net of \( \mathcal{F}_{\text{com}} \) with respect to \( \| \cdot \|_n \). Hence the covering number \( \mathcal{N}(u, \mathcal{F}_{\text{com}}; \| \cdot \|_n) \) is upper bounded by the cardinality of \( \hat{\mathcal{F}}_{\text{com}} \), 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 \( \int_0^1 H^{1/2}(u, \mathcal{F}_{\text{com}}; \| \cdot \|_n) \) can be upper bounded by

\[
\int_0^1 \{ \log \mathcal{N}(u/3, \mathcal{F}, \| \cdot \|_n) + \log \mathcal{N}(u/3, \mathcal{G}, \| \cdot \|_n) + \log \mathcal{N}(u/3, \mathcal{H}, \| \cdot \|_n) \}^{1/2} \, du
\leq \sqrt{V(\mathcal{F}) + V(\mathcal{G}) + V(\mathcal{H})} \int_0^1 \left\{ \sqrt{2 + \log(16C_{\text{VC}})} + \sqrt{2 \log(3u^{-1})} \right\} \, du.
\]

The desired result follows by applying Corollary 81 to the class \( \mathcal{F}_{\text{com}}, \) with \( \Psi_n(\mathcal{F}_{\text{com}}) \) taken to be the right-hand side of the preceding display. \( \blacksquare \)

References


