# Bregman contrastive estimation as a general framework for learning unnormalized models: Efficiency, robustness and optimal noise

**Yuto Fujii**[1] **Hiroaki Sasaki**[2] **Takafumi Kanamori**[1,3]

[1]Dept. of Math. & Comp. Sci., Institute of Science Tokyo, Tokyo, Japan
[2]Dept. of Math. Info., Meiji Gakuin University, Kanagawa, Japan
[3]RIKEN, Center for Advanced Intelligence Project, Tokyo, Japan

## Abstract

Learning unnormalized models is an ubiquitous problem in modern statistics and machine learning. Noise contrastive estimation (NCE) takes a popular approach based on solving a binary classification problem: Unnormalized models are learned by discriminating input data with artificially generated noise data. This paper follows this approach, but proposes a general framework based on the Bregman divergence. By selecting the convex function in the divergence, our framework includes existing methods as special cases, and novel variants of NCE can be also derived. Learning unnormalized models in the proposed framework is performed by estimating the posterior probability in the binary classification, i.e., the composition of the logistic function and the ratio of data and noise distributions. Due to the boundedness of the logistic function, our framework is advantageous in outlier-robustness. In fact, we theoretically prove that robust estimation is possible in our framework under a variety of convex functions in the Bregman divergence. Furthermore, the asymptotic efficiency is also investigated, implying a trade-off between efficiency and robustness in our framework. Inspired by these results, we derive the optimal noise distribution under some constraints for robustness. Finally, we numerically demonstrate the robustness of the novel variants of NCE and the derived optimal noise distribution.

## 1 INTRODUCTION

It is the central problem to learn unnormalized parametric models in modern statistics and machine learning. In contrast with the baseline methods (e.g., maximum likelihood estimation (MLE)), learning unnormalized models does not require them to be properly normalized, and thus offers a wider range of practical applications. A well-known application is generative models, e.g., generative adversarial networks (GAN) [Goodfellow et al., 2014] and diffusion models [Song and Ermon, 2019, Song et al., 2021]. Another example is self-supervised learning [Liu et al., 2020], which can be regarded as obtaining a useful representation of data through estimation of unnormalized models. Other applications include large language models [Mnih and Teh, 2012, Chen et al., 2024].

A number of practical methods for learning unnormalized models have been proposed so far [Besag, 1975, Hinton, 2002, Hyvärinen, 2005, Gutmann and Hyvärinen, 2012, Takenouchi and Kanamori, 2017, Uehara et al., 2020]. One of the promising methods is *noise contrastive estimation* (NCE) [Gutmann and Hyvärinen, 2012], which forms the basis of GAN and self-supervised learning. NCE learns unnormalized models by discriminating input data from artificially generated noise data through logistic regression, and thus can be seen as solving a binary classification problem. As in MLE [Van der Vaart, 1998, Wasserman, 2004], NCE has a good theoretical property called asymptotic efficiency [Gutmann and Hyvärinen, 2012]. On the other hand, this efficiency of NCE comes at the cost of robustness because logistic regression used in NCE is well-known to be sensitive to the contamination of outliers [Feng et al., 2014].

The crucial point in NCE is how to make noise data samples or choose the noise distribution because the practical performance strongly depends on them. To cope with this problem, Ceylan and Gutmann [2018] proposed to generate noise data by injecting a small amount of additive noise into input data, and demonstrated that the proposed approach works well for data with a lower-dimensional manifold structure (e.g., ring). GAN takes an iterative approach that updates noise samples (i.e., noise distribution) so that it makes difficult or almost impossible to discriminate them from input data samples Goodfellow et al. [2014]. A similar iterative approach has been proposed based on normalizing flow [Gao et al., 2020] where the noise distribution is

*Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence* (UAI 2026), PMLR 337:1604–1627.

updated towards the data distribution. See a recent review paper for NCE [Gutmann et al., 2022] and references therein. Overall, a common belief behind these approaches would be that the data or near-data distribution is an optimal choice for the noise distribution. However, in spite of this belief, it has recently been shown that the optimal noise distribution could be different from the data distribution [Chehab et al., 2022], meaning that there exists a clear gap between them.

This paper follows the same classification approach as NCE yet proposes a general framework to learn unnormalized models with the Bregman divergence [Bregman, 1967]. Previously, similar frameworks have been proposed based on the Bregman divergence, which directly estimate the ratio between data and noise distributions for learning unnormalized models [Gutmann and Hirayama, 2011, Uehara et al., 2020]. In contrast, the proposed framework estimates the posterior probability in the classification problem, i.e., the composition of the logistic function and ratio of the two probability distributions, and thus *not* directly the ratio. This subtle difference makes a great benefit in terms of robustness against outliers. The posterior probability is always bounded even if the ratio is unbounded due to the logistic function. This boundedness is critical for robust estimation because the ratio can be arbitrarily large when outliers move to tails of the data or noise distribution. Previously, an outlier-robust variant of NCE has been proposed [Sasaki and Takenouchi, 2024] but is also a special case of the proposed framework.

After proposing our framework, we perform detailed asymptotic analysis in robustness and efficiency. Our analysis based on the influence function [Huber and Ronchetti, 2009] guarantees the robustness of the proposed framework on a variety of convex functions in the Bregman divergence, while NCE is a rare exception, i.e., it is strongly hampered by the contamination by outliers. The asymptotic efficiency is achievable in the proposed framework as well with a particular choice of the noise distribution in contrast with NCE. Overall, these theoretical results imply a trade-off between efficiency and robustness in our framework.

Subsequently, we establish the optimal noise distribution in the proposed framework. In stark contrast with Chehab et al. [2022], we propose to impose novel constraints to optimize the noise distribution. Interestingly, the optimal noise distribution derived under one of the constraints is a combination of the noise distribution in Chehab et al. [2022] and data distribution, and thus might give an insight to bridge a gap to the aforementioned common belief behind NCE. Furthermore, inspired by the asymptotic analysis, we propose other constraints for robustness, and derive the optimal noise distributions under them with substantially different proofs from that in Chehab et al. [2022]. Finally, we experimentally show the robustness of the derived optimal noise distribution and novel variants of NCE in the proposed framework.

This paper is organized as follows: Section 2 formulates the problem in this paper and reviews noise contrastive estimation and existing frameworks for learning unnormalized models. Our framework based on the Bregman divergence is proposed in Section 3. Section 4 performs asymptotic analysis in terms of efficiency and robustness, while Section 5 derives the optimal noise distribution in the proposed framework. The novel variants of NCE in our framework and the optimal noise distribution are experimentally investigated in Section 6. Section 7 concludes this paper. A notation table is given in Appendix A.

## 2 BACKGROUND

We first suppose that there exists a $\sigma$-finite reference measure space $(\Omega, \mathcal{B}, \mu)$, and the probability measure $P_{\mathrm{d}}$ for data is absolutely continuous with respect to $\mu$, i.e., it has the density as $p_{\mathrm{d}} = \frac{\mathrm{d}P_{\mathrm{d}}}{\mathrm{d}\mu}$. Then, we assume that $n$ samples of data vectors $\boldsymbol{x}_i$ drawn from $p_{\mathrm{d}}(\boldsymbol{x})$ are available as

$$\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{iid}}{\sim} p_{\mathrm{d}}(\boldsymbol{x}).$$

Then, this paper considers the problem of estimating the parameter vector $\boldsymbol{\theta}$ in a parametric model,

$$p_{\mathrm{m}}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{q_{\mathrm{m}}(\boldsymbol{x}; \boldsymbol{\theta})}{Z(\boldsymbol{\theta})},$$

from $\mathcal{X}$ where $Z(\boldsymbol{\theta})$ is the partition function, and $q_{\mathrm{m}}(\boldsymbol{x}; \boldsymbol{\theta})$ is some non-negative function with $\boldsymbol{\theta}$. This section reviews noise contrastive estimation and an existing framework for learning unnormalized models based on the Bregman divergence.

### 2.1 NOISE CONTRASTIVE ESTIMATION

Noise contrastive estimation (NCE) learns a parametric model by solving a binary classification problem where the following dataset is supposed to be additionally available:

$$\mathcal{Y} = \{\boldsymbol{y}_i\}_{i=1}^{n_{\mathrm{y}}} \overset{\text{iid}}{\sim} p_{\mathrm{n}}(\boldsymbol{y}),$$

where $\boldsymbol{y}_i$ and $p_{\mathrm{n}}$ are called the *noise samples* and *noise distribution*, respectively[1]. $p_{\mathrm{n}}$ is fundamentally required to have two properties: (i) the functional form of $p_{\mathrm{n}}$ is known and (ii) sampling from $p_{\mathrm{n}}$ is possible. By regarding $\mathcal{X}$ and $\mathcal{Y}$ as datasets from binary classes, the conditional distribution of data $\boldsymbol{u}$ given class label $C$ is defined as

$$p(\boldsymbol{u}|C=0) = p_{\mathrm{d}}(\boldsymbol{u}) \quad \text{and} \quad p(\boldsymbol{u}|C=1) = p_{\mathrm{n}}(\boldsymbol{u}).$$

[1] As in $p_{\mathrm{d}}$, we assume that the probability measure $P_{\mathrm{n}}$ for noise data is absolutely continuous with respect to $\mu$ and has the density as $p_{\mathrm{n}} = \frac{\mathrm{d}P_{\mathrm{n}}}{\mathrm{d}\mu}$.

Then, the posterior probability is formulated as

$$p(C=0|\boldsymbol{u}) = \frac{p_{\mathrm{d}}(\boldsymbol{u})}{p_{\mathrm{d}}(\boldsymbol{u}) + \nu p_{\mathrm{n}}(\boldsymbol{u})} = \frac{r(\boldsymbol{u})}{r(\boldsymbol{u}) + \nu}$$
$$p(C=1|\boldsymbol{u}) = \frac{\nu p_{\mathrm{n}}(\boldsymbol{u})}{p_{\mathrm{d}}(\boldsymbol{u}) + \nu p_{\mathrm{n}}(\boldsymbol{u})} = \frac{\nu}{r(\boldsymbol{u}) + \nu}, \quad (1)$$

where $\nu := \frac{p(C=1)}{p(C=0)}$ with the class probability $p(C)$ and $r(\boldsymbol{u}) := p_{\mathrm{d}}(\boldsymbol{u})/p_{\mathrm{n}}(\boldsymbol{u})$ is the ratio of two probability distributions.

Learning in NCE is performed through the posterior estimation. To this end, NCE expresses the partition function as a single parameter $c$ and defines an unnormalized model $p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha})$ as

$$\log p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}) = \log q_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\theta}) - c.$$

Then, all parameters $\boldsymbol{\alpha} = (\boldsymbol{\theta}, c)$ are estimated by fitting the following model to the true posterior probability:

$$f_0(\boldsymbol{u};\boldsymbol{\alpha}) = \frac{p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha})}{p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}) + \nu p_{\mathrm{n}}(\boldsymbol{u})} = \frac{r_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha})}{r_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}) + \nu}$$
$$f_1(\boldsymbol{u};\boldsymbol{\alpha}) = \frac{\nu p_{\mathrm{n}}(\boldsymbol{u})}{p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}) + \nu p_{\mathrm{n}}(\boldsymbol{u})} = \frac{\nu}{r_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}) + \nu}, \quad (2)$$

where $r_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}) := p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha})/p_{\mathrm{n}}(\boldsymbol{u})$. Then, the risk function of NCE is formulated through logistic regression as

$$\mathcal{L}_{\mathrm{lr}}(\boldsymbol{\alpha}) = -\mathbb{E}_{\mathrm{d}}[\log f_0(\boldsymbol{X};\boldsymbol{\alpha})] - \nu\mathbb{E}_{\mathrm{n}}[\log f_1(\boldsymbol{Y};\boldsymbol{\alpha})],$$

where $\mathbb{E}_{\mathrm{d}}$ and $\mathbb{E}_{\mathrm{n}}$ denote the expectations over $p_{\mathrm{d}}$ and $p_{\mathrm{n}}$, respectively. In practice, the empirical version of $\mathcal{L}_{\mathrm{lr}}(\boldsymbol{\alpha})$ is minimized to estimate the parameter vector $\boldsymbol{\alpha}$.

### 2.2 DIRECT RATIO ESTIMATION WITH BREGMAN DIVERGENCE

Learning unnormalized models is possible by estimating the ratio $r_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}) = p^0_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})/p_{\mathrm{n}}(\boldsymbol{x})$ if $p_{\mathrm{n}}$ is known as assumed in NCE. In fact, (2) indicates that NCE can be regarded as estimating $r_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})$ through logistic regression. Based on the ratio estimation, a general framework for learning unnormalized models has been proposed [Gutmann and Hirayama, 2011, Uehara et al., 2020] with the (separable) Bregman divergence, whose risk is defined by

$$-\mathbb{E}_{\mathrm{d}}\left[\varphi'\left(r_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha})\right)\right] + \mathbb{E}_{\mathrm{n}}\left[\left\{\varphi'\left(r_{\mathrm{m}}(\boldsymbol{Y};\boldsymbol{\alpha})\right) r_{\mathrm{m}}(\boldsymbol{Y};\boldsymbol{\alpha}) - \varphi\left(r_{\mathrm{m}}(\boldsymbol{Y};\boldsymbol{\alpha})\right)\right],$$

where $\varphi(\cdot)$ is a strictly convex function on $\mathbb{R}_+$ and $\varphi'(z) := \frac{\mathrm{d}}{\mathrm{d}z}\varphi(z)$. By selecting $\varphi$, a variety of risk functions can be derived.

This line of work is similar to our approach, but there exists a fundamental difference: We do not directly estimate the ratio $r(\boldsymbol{x})$, but the posterior probability in (1), i.e., the composition of the logistic function and ratio $r(\boldsymbol{x})$. The notable point is that the posterior probability is always bounded due to the composition by the logistic function, while ratio $r(\boldsymbol{x})$ can be unbounded. This subtle point makes robust learning possible when outliers lie on the tails of the data or noise distribution. Indeed, it has been theoretically shown that direct ratio estimation can be strongly hampered by the contamination of outliers even when robust divergences are employed [Kanamori and Sugiyama, 2014, Sasaki and Takenouchi, 2024].

## 3 BREGMAN CONTRASTIVE ESTIMATION

The main idea in our framework is to employ the Bregman divergence between the posterior probability $p(C|\boldsymbol{u})$ and its model $f_C(\boldsymbol{u};\boldsymbol{\alpha})$ for which the risk function is defined as

$$\int p(\boldsymbol{u}) \sum_{C\in\{0,1\}} [\varphi(p(C|\boldsymbol{u})) - \varphi(f_C(\boldsymbol{u};\boldsymbol{\alpha})) - \varphi'(f_C(\boldsymbol{u};\boldsymbol{\alpha}))(p(C|\boldsymbol{u}) - f_C(\boldsymbol{u};\boldsymbol{\alpha}))]\,\mathrm{d}\boldsymbol{u}. \quad (3)$$

Then, by expressing the marginal distribution as

$$p(\boldsymbol{u}) = \sum_{C\in\{0,1\}} p(\boldsymbol{u}|C)p(C) \propto p_{\mathrm{d}}(\boldsymbol{u}) + \nu p_{\mathrm{n}}(\boldsymbol{u}),$$

Appendix B shows that our risk function can be written by

$$\mathcal{L}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha}) := -\mathbb{E}_{\mathrm{d}}\left[\Phi_{\varphi}(\boldsymbol{X};\boldsymbol{\alpha}) + \Psi_{\varphi}(\boldsymbol{X};\boldsymbol{\alpha}) f_1(\boldsymbol{X};\boldsymbol{\alpha})\right] - \nu\mathbb{E}_{\mathrm{n}}\left[\Phi_{\varphi}(\boldsymbol{Y};\boldsymbol{\alpha}) - \Psi_{\varphi}(\boldsymbol{Y};\boldsymbol{\alpha}) f_0(\boldsymbol{Y};\boldsymbol{\alpha})\right], \quad (4)$$

where $\Phi_{\varphi}(\boldsymbol{u};\boldsymbol{\alpha}) := \varphi\left(f_0(\boldsymbol{u};\boldsymbol{\alpha})\right) + \varphi\left(f_1(\boldsymbol{u};\boldsymbol{\alpha})\right)$ and $\Psi_{\varphi}(\boldsymbol{u};\boldsymbol{\alpha}) := \varphi'\left(f_0(\boldsymbol{u};\boldsymbol{\alpha})\right) - \varphi'\left(f_1(\boldsymbol{u};\boldsymbol{\alpha})\right)$. In practice, $\mathcal{L}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha})$ is empirically approximated from data samples as

$$\widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha}) := -\frac{1}{n}\sum_{i=1}^{n}\left[\Phi_{\varphi}(\boldsymbol{x}_i;\boldsymbol{\alpha}) + \Psi_{\varphi}(\boldsymbol{x}_i;\boldsymbol{\alpha}) f_1(\boldsymbol{x}_i;\boldsymbol{\alpha})\right] - \frac{1}{n}\sum_{i=1}^{n_{\mathrm{y}}}\left[\Phi_{\varphi}(\boldsymbol{y}_i;\boldsymbol{\alpha}) - \Psi_{\varphi}(\boldsymbol{y}_i;\boldsymbol{\alpha}) f_0(\boldsymbol{y}_i;\boldsymbol{\alpha})\right],$$

where $\nu$ is substituted by $\hat{\nu} := n_{\mathrm{y}}/n$. The parameter vector $\boldsymbol{\alpha}$ is estimated by minimizing $\widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha})$ as

$$\widehat{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha}) = (\widehat{\boldsymbol{\theta}}, \widehat{c}).$$

We call this method as the *Bregman contrastive estimation* (BCE).

Next, we show that BCE includes NCE and its robust variant called the power NCE (PNCE) [Sasaki and Takenouchi, 2024] as special cases:

**Example 1** (NCE)**.** *When $\varphi(z) = z\log z$, $\mathcal{L}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha})$ is reduced to $\mathcal{L}_{\mathrm{lr}}(\boldsymbol{\alpha})$.*

Table 1: Examples of BCE under the choices of $\varphi$.

| | $\varphi(z)$ | risk | reference |
|---|---|---|---|
| NCE | $z \log z$ | $\mathcal{L}_{\mathrm{lr}}(\boldsymbol{\alpha})$ | [Gutmann and Hyvärinen, 2012] |
| PNCE | $\frac{z^{\beta+1}}{\beta(\beta+1)}$ | $\mathcal{L}_{\beta}(\boldsymbol{\alpha})$ | [Sasaki and Takenouchi, 2024] |
| $\rho$-NCE | $-\log(z+\rho)$ | $\mathcal{L}_{\rho}(\boldsymbol{\alpha})$ | **This work** |
| $\tau$-NCE | $(z+\tau)\log\frac{z+\tau}{1+\tau}$ | $\mathcal{L}_{\tau}(\boldsymbol{\alpha})$ | **This work** |

**Example 2** (PNCE). *When $\varphi(z) = z^{\beta+1}/(\beta(\beta+1))$, $\mathcal{L}_{\varphi}^{\mathrm{SB}}(\boldsymbol{\alpha})$ is reduced to the following risk function proposed in Sasaki and Takenouchi [2024]:*

$$\mathcal{L}_{\beta}(\boldsymbol{\alpha}) = -\frac{1}{\beta}\left(\mathbb{E}_{\mathrm{d}}[f_0(\boldsymbol{X};\boldsymbol{\alpha})^{\beta}] + \nu\mathbb{E}_{\mathrm{n}}[f_1(\boldsymbol{Y};\boldsymbol{\alpha})^{\beta}]\right)$$
$$+\frac{1}{\beta+1}\sum_{c=0}^{1}(\mathbb{E}_{\mathrm{d}}[f_c(\boldsymbol{X};\boldsymbol{\alpha})^{\beta+1}] + \nu\mathbb{E}_{\mathrm{n}}[f_c(\boldsymbol{X};\boldsymbol{\alpha})^{\beta+1}]).$$

Next, two novel variants of NCE are derived from BCE. The first choice is $\varphi(z) = -\log(z+\rho)$, which has been used in boosting [Murata et al., 2004] and introduced to cope with a mislabeling problem [Takenouchi and Eguchi, 2004]. The second choice can be seen as a generalization of the first one.

**Example 3** ($\rho$-NCE). *When $\varphi(z) = -\log(z+\rho)$ with a parameter $\rho > 0$, $\mathcal{L}_{\varphi}^{\mathrm{SB}}(\boldsymbol{\alpha})$ is reduced to*

$$\mathcal{L}_{\rho}(\boldsymbol{\alpha}) = \mathbb{E}_{\mathrm{d}}\Bigg[-\log\left(f_0(\boldsymbol{X};\boldsymbol{\alpha})f_1(\boldsymbol{X};\boldsymbol{\alpha}) + \rho^2 + \rho\right) - \frac{f_1(\boldsymbol{X};\boldsymbol{\alpha})\{f_0(\boldsymbol{X};\boldsymbol{\alpha}) - f_1(\boldsymbol{X};\boldsymbol{\alpha})\}}{f_0(\boldsymbol{X};\boldsymbol{\alpha})f_1(\boldsymbol{X};\boldsymbol{\alpha}) + \rho^2 + \rho}\Bigg]$$
$$-\nu\mathbb{E}_{\mathrm{n}}\Bigg[-\log\left(f_0(\boldsymbol{Y};\boldsymbol{\alpha})f_1(\boldsymbol{Y};\boldsymbol{\alpha}) + \rho^2 + \rho\right) - \frac{f_0(\boldsymbol{Y};\boldsymbol{\alpha})\{f_0(\boldsymbol{Y};\boldsymbol{\alpha}) - f_1(\boldsymbol{Y};\boldsymbol{\alpha})\}}{f_0(\boldsymbol{Y};\boldsymbol{\alpha})f_1(\boldsymbol{Y};\boldsymbol{\alpha}) + \rho^2 + \rho}\Bigg].$$

*We call the learning method based on $\mathcal{L}_{\rho}(\boldsymbol{\alpha})$ as $\rho$-NCE.*

**Example 4** ($\tau$-NCE). *When $\varphi(z) = (z+\tau)\log\frac{z+\tau}{1+\tau}$ for $\tau > 0$, $\mathcal{L}_{\varphi}^{\mathrm{SB}}(\boldsymbol{\alpha})$ is reduced to*

$$\mathcal{L}_{\tau}(\boldsymbol{\alpha}) = -\mathbb{E}_{\mathrm{d}}\left[(1+\tau)\log f_0^{\tau}(\boldsymbol{X};\boldsymbol{\alpha}) + \tau\log f_1^{\tau}(\boldsymbol{X};\boldsymbol{\alpha})\right] - \nu\mathbb{E}_{\mathrm{n}}\left[(1+\tau)\log f_1^{\tau}(\boldsymbol{Y};\boldsymbol{\alpha}) + \tau\log f_0^{\tau}(\boldsymbol{Y};\boldsymbol{\alpha})\right],$$

*where $f_C^{\tau}(\boldsymbol{x};\boldsymbol{\alpha}) := \frac{f_C(\boldsymbol{x};\boldsymbol{\alpha})+\tau}{1+\tau}$ for $C = 0, 1$. We call the learning method based on $\mathcal{L}_{\tau}(\boldsymbol{\alpha})$ as $\tau$-NCE.*

It is shown theoretically and empirically that $\tau$-NCE and $\rho$-NCE are robust against outliers. Table 1 summarizes how each method corresponds to a particular choice of $\varphi$.

### 3.1 MODEL SELECTION IN BCE

The selection of the hyper-parameters (e.g., $\rho$ and $\tau$) is important in practice. Here, by following a simple heuristic approach proposed in Minami and Eguchi [2003] and Sasaki and Takenouchi [2024], a method for model selection based on cross-validation is presented as follows: Taking $\tau$-NCE as an example,

1. Choose a fixed parameter $\tau_0$ called the *anchor parameter* and candidates of $\tau$ denoted by $T$.
2. Divide $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ into $K$ disjoint subsets $\{\mathcal{X}_k\}_{k=1}^{K}$.
3. Obtain the estimator $\widehat{\boldsymbol{\alpha}}_{\tau}^{(k)}$ for $\tau \in T$ from $\mathcal{X}$ without $\mathcal{X}_k$ and reference data vectors $\boldsymbol{y}_j$ which are generated by keeping the same sample ratio $\hat{\nu}$.
4. Compute $\widehat{\mathcal{L}}_{\tau_0}$ with the anchor parameter $\tau_0$ from $\mathcal{X}_k$ as $\mathrm{CV}(k,\tau) := \widehat{\mathcal{L}}_{\tau_0}(\widehat{\boldsymbol{\alpha}}_{\tau}^{(k)})$.
5. Choose the optimal value of $\tau$ on $T$ that minimizes $\frac{1}{K}\sum_{k=1}^{K}\mathrm{CV}(k,\tau)$.

It seems to still matter the choice of the anchor parameter in cross-validation, but our empirical results show that this method reasonably works well on a range of the anchor parameters.

## 4 ASYMPTOTIC ANALYSIS: ROBUSTNESS AND EFFICIENCY

This section performs asymptotic analysis to BCE in terms of robustness and efficiency. Throughout this section, let us assume that for all $\boldsymbol{\theta}$,

$$\mathcal{D} := \mathrm{supp}(p_{\mathrm{m}}(\cdot;\boldsymbol{\theta})) \subseteq \mathrm{supp}(p_{\mathrm{n}}), \tag{5}$$

where $\mathrm{supp}(p_{\mathrm{m}}(\cdot;\boldsymbol{\theta})) := \{\boldsymbol{x} \mid p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta}) \neq 0\}$ and $\mathrm{supp}(p_{\mathrm{n}}) := \{\boldsymbol{y} \mid p_{\mathrm{n}}(\boldsymbol{y}) \neq 0\}$. This support assumption is standard in the analysis of NCE [Gutmann and Hyvärinen, 2012, Matsuda et al., 2021, Sasaki and Takenouchi, 2024], and required to estimate $p_{\mathrm{m}}(\cdot;\boldsymbol{\theta})$ over the full support. Furthermore, we assume that $p_{\mathrm{m}}^{0}(\boldsymbol{x};\boldsymbol{\alpha})$ is differentiable with respect to $\boldsymbol{\alpha}$ such that the gradient $\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}) := \nabla_{\boldsymbol{\alpha}}\log p_{\mathrm{m}}^{0}(\boldsymbol{x};\boldsymbol{\alpha})$ is properly computed.

### 4.1 ROBUSTNESS OF BCE

Here, the robustness of BCE is investigated based on the *influence function* [Huber and Ronchetti, 2009, Hampel et al., 2011, Maronna et al., 2019]. To this end, we start by

modeling the contaminated probability distributions as

$$\bar{p}_{\mathrm{d}}(\boldsymbol{x}) = (1-\varepsilon)p_{\mathrm{d}}(\boldsymbol{x}) + \varepsilon\delta_{\boldsymbol{x}_{\mathrm{o}}}(\boldsymbol{x}), \tag{6}$$

where $0 \leq \varepsilon < 1$ is a constant, $\boldsymbol{x}_{\mathrm{o}}$ is an outlier, and $\delta_{\boldsymbol{x}_{\mathrm{o}}}(\cdot)$ denotes the Dirac measure centered at $\boldsymbol{x}_{\mathrm{o}}$ and is sufficient for the worse-case local analysis. As the contamination ratio $\varepsilon$ is increased, the dataset is contaminated by more outliers. Based on the contaminated probability distribution (6), we define the following estimators:

$$\boldsymbol{\alpha}_\star = \operatorname*{argmin}_{\boldsymbol{\alpha}} \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) \text{ and } \boldsymbol{\alpha}_\varepsilon = \operatorname*{argmin}_{\boldsymbol{\alpha}} \bar{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}), \tag{7}$$

where by expressing the expectation over $\bar{p}_{\mathrm{d}}(\cdot)$ as $\bar{\mathbb{E}}_{\mathrm{d}}$,

$$\begin{aligned}\bar{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) := &-\bar{\mathbb{E}}_{\mathrm{d}}\left[\Phi(\boldsymbol{X};\boldsymbol{\alpha}) + \Psi(\boldsymbol{X};\boldsymbol{\alpha})f_1(\boldsymbol{X};\boldsymbol{\alpha})\right] \\ &- \nu\mathbb{E}_{\mathrm{n}}\left[\Phi(\boldsymbol{Y};\boldsymbol{\alpha}) - \Psi(\boldsymbol{Y};\boldsymbol{\alpha})f_0(\boldsymbol{Y};\boldsymbol{\alpha})\right].\end{aligned}$$

Based on these estimators, the influence function (IF) is defined in the limit of $\varepsilon \to 0$ as follows:

$$\mathrm{IF}(\boldsymbol{x}_{\mathrm{o}}) = \lim_{\varepsilon\to 0} \frac{\boldsymbol{\alpha}_\varepsilon - \boldsymbol{\alpha}_\star}{\varepsilon}. \tag{8}$$

$\mathrm{IF}(\boldsymbol{x}_{\mathrm{o}})$ helps us to understand how the asymptotic estimator $\boldsymbol{\alpha}_\star$ is influenced by infinitesimal contamination of outliers. A useful notion based on IF is *B-robustness*: If $\sup_{\boldsymbol{x}\in\mathcal{D}} \|\mathrm{IF}(\boldsymbol{x})\| < \infty$, $\boldsymbol{\alpha}_\star$ is said to be B-robust. B-robustness ensures that the influence from outliers $\boldsymbol{x}_{\mathrm{o}}$ is limited even if they go far away from the center of data.

We first derive IF and show a condition for B-robustness in the following proposition, and then discuss how mild the condition is:

**Proposition 1.** *Assume that the Hessian matrix $\boldsymbol{H}_\varphi = \nabla^2_{\boldsymbol{\alpha}}\mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_\star}$ is invertible. Then, the influence function is given by*

$$\begin{aligned}\mathrm{IF}(\boldsymbol{x}_{\mathrm{o}}) = \boldsymbol{H}_\varphi^{-1}\{&\mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)f_1(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)] \\ &- w_\varphi(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha}_\star)f_1(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha}_\star)\},\end{aligned} \tag{9}$$

*where with $\varphi''(z) := \frac{\mathrm{d}^2}{\mathrm{d}z^2}\varphi(z)$,*

$$\begin{aligned}&w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}) \\ &:= \{\varphi''(f_0(\boldsymbol{x};\boldsymbol{\alpha})) + \varphi''(f_1(\boldsymbol{x};\boldsymbol{\alpha}))\}\, f_0(\boldsymbol{x};\boldsymbol{\alpha})f_1(\boldsymbol{x};\boldsymbol{\alpha}).\end{aligned} \tag{10}$$

*Furthermore, $\boldsymbol{\alpha}_\star$ is B-robust if the following inequality holds:*

$$\sup_{\boldsymbol{x}\in\mathcal{D}} w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star)\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)\| < \infty. \tag{11}$$

The proof is deferred to Appendix C. The first term on the right-hand side of (9) is a constant bias and cannot be eliminated but has no influence from outliers. The second term is influenced by outlier $\boldsymbol{x}_{\mathrm{o}}$, and thus $\boldsymbol{\alpha}_\star$ is guaranteed to be B-robust if this term is finite, i.e., condition (11) is fulfilled.

For NCE, it is very unlikely that condition (11) is satisfied because the choice of $\varphi(z) = z\log z$ in Example 1 yields a constant weight $w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star) = 1$. Thus, the left-hand side on (11) becomes $\sup_{\boldsymbol{x}\in\mathcal{D}} \|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)\|$, which can be unbounded in general. Non-robustness of NCE has been previously reported in Sasaki and Takenouchi [2024] as well. Appendix D illustrates that weights $w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star)$ for PNCE, $\tau$-NCE and $\rho$-NCE are non-constant and approach to zero if $f_0 \to 0$ or $f_1 \to 1$.

Next, we show when condition (11) is fulfilled. To enhance the interpretability, we re-express condition (11) in terms of the ratio model $r_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}) = p^0_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})/p_{\mathrm{n}}(\boldsymbol{x};\boldsymbol{\alpha})$ in the following corollary where $\mathcal{C}^k$ denotes the space of continuous and $k$-times differentiable functions:

**Corollary 1.** *Assume that $w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star)$ is not a constant function and $\varphi$ belongs to $\mathcal{C}^2$. Then,* (11) *is equivalent to*

$$\sup_{\boldsymbol{x}\in\mathcal{D}} \frac{\nu r_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)}{(\nu + r_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star))^2}\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)\| < \infty. \tag{12}$$

The proof is a direct consequence from the finiteness of $\varphi''$ because $\varphi$ is a $\mathcal{C}^2$ function and the domain of $\varphi$ is $[0,1]$. It is useful to consider a common situation (e.g., Gaussian distributions) where $\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)\|$ diverges when $p^0_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star) \to 0$ as $\boldsymbol{x}$ goes to the boundary of $\mathcal{D}$. In this situation, $r_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})$ approaches zero if $p_{\mathrm{n}}(\boldsymbol{x})$ has a wider support and slower decay than $p^0_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)$ because the growth speed of $\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star) = \nabla_{\boldsymbol{\alpha}} \log p^0_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_\star}$ is logarithmic. Thus, condition (12) seems very mild and could be satisfied under a variety of strictly convex functions $\varphi$ because the fundamental condition is that $\varphi$ is a $\mathcal{C}^2$ function. Moreover, a large value of $\nu$ would also make condition (12) more easily satisfied, and thus improve the robustness of BCE.

### 4.2 EFFICIENCY OF BCE

Efficiency of statistical estimators is a classical yet important notion. An estimator is said to be *asymptotically efficient* if the asymptotic variance is equal to the Cramér-Rao lower bound [Lehmann and Casella, 1998], which is the inverse of the Fisher information matrix[2] defined as

$$\boldsymbol{I}_{\mathrm{F}} := \mathbb{E}_{\mathrm{d}}\left[\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right].$$

To investigate the efficiency, we first establish the asymptotic normality of BCE and then show some differences to NCE. The consistency of BCE is proved in Appendix E.

[2]More formally, the Fisher information matrix $\boldsymbol{I}_{\mathrm{F}}$ is defined with the expectation over $p_{\mathrm{m}}(\cdot,\boldsymbol{\theta})$. However, Assumptions (A3-4) imply $p_{\mathrm{m}}(\cdot;\boldsymbol{\theta}_\star) = p_{\mathrm{d}}(\cdot)$. Therefore, we defined $\boldsymbol{I}_{\mathrm{F}}$ in terms of the expectation over $p_{\mathrm{d}}(\cdot)$.

For asymptotic normality, we make the following regularity assumptions:

(A1) The parameter space $\Theta$ of $\boldsymbol{\theta}$ is compact.

(A2) For all $\boldsymbol{\theta}$, $\operatorname{supp} p_{\mathrm{m}}(\cdot;\boldsymbol{\theta}) \subset \operatorname{supp} p_{\mathrm{n}}$.

(A3) $p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta})$ is identifiable, i.e., $p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta}_1) = p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta}_2)$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

(A4) $p_{\mathrm{d}}(\boldsymbol{x})$ belongs to the same parametric family as $p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta})$, i.e., there exists a constant vector $\boldsymbol{\theta}_0 \in \mathrm{Int}(\Theta)$ (i.e., interior of $\Theta$) such that $p_{\mathrm{d}}(\boldsymbol{x}) = p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta}_0)$.

(A5) For all $\epsilon > 0$, $\inf_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_\star\|\geq\epsilon} \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) > \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star)$.

(A6) $\sup_{\boldsymbol{\alpha}} |\mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) - \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})| \to 0$ in probability.

Assumptions (A1-4) are quite standard. Assumption (A5) means that $\boldsymbol{\alpha}_\star$ is a well-separated point of minimum of $\mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})$ [Van der Vaart, 1998], and Assumption (A6) is the well-known uniform law of large numbers. By definition of $\boldsymbol{\alpha}_\star$ and Assumption (A4), we have $p_{\mathrm{d}}(\boldsymbol{x}) = p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta}_\star)$. Therefore, the identifiability condition (A3) ensures that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_\star$, which enables us to express $f_1(\boldsymbol{u};\boldsymbol{\alpha}_\star)$ as

$$f_1^\star(\boldsymbol{u}) := f_1(\boldsymbol{u};\boldsymbol{\alpha}_\star) = \frac{\nu p_{\mathrm{n}}(\boldsymbol{u})}{p_{\mathrm{d}}(\boldsymbol{u}) + \nu p_{\mathrm{n}}(\boldsymbol{u})}. \tag{13}$$

In addition, as in Gutmann and Hyvärinen [2012], Matsuda et al. [2021], we suppose the asymptotics that $\hat{\nu} = n_{\mathrm{y}}/n \to \nu$ as $n \to \infty$ and $n_{\mathrm{y}} \to \infty$.

We first establish the asymptotic normality of BCE in the following theorem proved in Appendix F:

**Theorem 1** (Asymptotic normality)**.** *Suppose that Assumptions (A1-6) hold. We further assume that there exists an integrable function $b(\boldsymbol{x})$ such that*

$$\left|\partial_{ij} p_{\mathrm{m}}^0(\boldsymbol{x};\boldsymbol{\alpha})\right| \leq b(\boldsymbol{x}) \text{ and } \left|\partial_{ijk} p_{\mathrm{m}}^0(\boldsymbol{x};\boldsymbol{\alpha})\right| \leq b(\boldsymbol{x}),$$

*where $\partial_{ij} := \frac{\partial^2}{\partial\alpha_i\partial\alpha_j}$ and $\partial_{ijk} := \frac{\partial^3}{\partial\alpha_i\partial\alpha_j\partial\alpha_k}$. Then, $\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star)$ converges to the following normal distribution:*

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Lambda}_\varphi),$$

*where $\boldsymbol{\Lambda}_\varphi := \boldsymbol{J}_\varphi^{-1}\boldsymbol{I}_\varphi\boldsymbol{J}_\varphi^{-1}$, and with $\boldsymbol{g}_{\mathrm{m}}^w(\boldsymbol{x};\boldsymbol{\alpha}) := w_\varphi(\boldsymbol{x};\boldsymbol{\alpha})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})$,*

$$\begin{aligned}\boldsymbol{I}_\varphi = \mathbb{E}_{\mathrm{d}}\left[f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}^w(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}^w(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right] - \left(1+\frac{1}{\nu}\right) \\ \times \mathbb{E}_{\mathrm{d}}\left[f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}^w(\boldsymbol{X};\boldsymbol{\alpha}_\star)\right]\mathbb{E}_{\mathrm{d}}\left[f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}^w(\boldsymbol{X};\boldsymbol{\alpha}_\star)\right]^\top\end{aligned} \tag{14}$$

$$\boldsymbol{J}_\varphi = \mathbb{E}_{\mathrm{d}}\left[f_1^\star(\boldsymbol{X})w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right]. \tag{15}$$

The asymptotic variance $\boldsymbol{\Lambda}_\varphi$ of BCE takes a complex form and clearly depends on the strictly convex function $\varphi$. The following theorem asserts that NCE is in fact the optimal choice of $\varphi$ in the sense that it is the closest to the Cramér-Rao bound:

**Theorem 2** (Optimality of NCE)**.** *For all strictly convex functions $\varphi$, it holds*

$$\boldsymbol{\Lambda}_\varphi \succeq \boldsymbol{\Lambda}_{\mathrm{NCE}}, \tag{16}$$

*where $\boldsymbol{\Lambda}_\varphi \succeq \boldsymbol{\Lambda}_{\mathrm{NCE}}$ means that $\boldsymbol{\Lambda}_\varphi - \boldsymbol{\Lambda}_{\mathrm{NCE}}$ is positive semidefinite, and $\boldsymbol{\Lambda}_{\mathrm{NCE}}$ denotes the asymptotic variance of NCE and is given by* (33)*. Equality in* (16) *holds when $w_\varphi$ is a constant function.*

The proof is deferred in Appendix G. Theorem 2 indicates the optimality of NCE with respect to the choice of $\varphi$. On the other hand, as shown in Section 4.1, NCE is not robust against outliers This may imply a trade-off in terms of the choice of $\varphi$: An efficient choice of $\varphi$ comes at a cost of loosing outlier-robustness.

NCE has been shown to achieve the Cramér-Rao bound in the limit of $\nu \to \infty$ regardless of the choice of the noise distribution $p_{\mathrm{n}}$ [Gutmann and Hyvärinen, 2012]. In contrast, the following theorem indicates that the Cramér-Rao bound is attained in BCE under a particular choice of $p_{\mathrm{n}}$ in the asymptotic regime as for $\nu$:

**Proposition 2** (Noise limit)**.** *Assume that $\varphi$ belongs to $\mathcal{C}^2$. Then, as $\nu \to \infty$,*

$$\boldsymbol{\Lambda}_\varphi = \tilde{\boldsymbol{J}}^{-1}\tilde{\boldsymbol{I}}\tilde{\boldsymbol{J}}^{-1} + O(\nu^{-2}),$$

*where*

$$\begin{aligned}\tilde{\boldsymbol{I}} &= \mathbb{E}_{\mathrm{d}}\left[\left(\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})}\right)^2 \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right] \\ &\quad - \mathbb{E}_{\mathrm{d}}\left[\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\right]\mathbb{E}_{\mathrm{d}}\left[\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\right]^\top \\ \tilde{\boldsymbol{J}} &= \mathbb{E}_{\mathrm{d}}\left[\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right].\end{aligned}$$

The proof is given in Appendix H. In the limit of $\nu \to \infty$, the striking features of $\boldsymbol{\Lambda}_\varphi$ are twofold: (i) $\boldsymbol{\Lambda}_\varphi$ does not depend on the choice of $\varphi$, and (ii) unlike NCE, the noise distribution $p_{\mathrm{n}}$ still exists in $\boldsymbol{\Lambda}_\varphi$. This possibly indicates that parameter estimation by BCE is more dependent on the choice of $p_{\mathrm{n}}$ than $\varphi$ in a large noise sample regime.

Proposition 2 shows that $\boldsymbol{\Lambda}_\varphi$ asymptotically equals to the inverse of $\boldsymbol{I}_{\mathrm{F}}$, i.e., BCE achieves the Cramér-Rao bound when $p_{\mathrm{n}} = p_{\mathrm{d}}$. However, this choice of $p_{\mathrm{n}}$ makes the weight function $w_\varphi$ in (10) to be constant, and thus condition (11) for B-robustness is never satisfied when $\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)\|$ diverges. This again may reveal a trade-off between efficiency and robustness, but this trade-off seems to be slightly different from the aforementioned one after Theorem 2: The choice of $p_{\mathrm{n}}$ is more relevant in the asymptotic regime (i.e., $\nu \to \infty$), while the choice of $\varphi$ would matter in non-asymptotic regime as implied in Theorem 2.

## 5 THE OPTIMAL NOISE DISTRIBUTION WITH ROBUSTNESS

A common criterion to measure the statistical performance is *mean-squared error* (MSE) between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_\star$. As in Chehab et al. [2022], we derive the optimal noise distribution based on the following approximation of MSE:

$$\mathrm{MSE} := \frac{\mathrm{Tr}(\boldsymbol{\Lambda}_\varphi)}{n}. \tag{17}$$

This approximation would be valid when $n$ is large because the bias term in MSE often more quickly goes to zero as $n$ increases. In this section, we express $\boldsymbol{g}_\mathrm{m}(\boldsymbol{x}) := \nabla_{\boldsymbol{\theta}} \log p_\mathrm{m}(\boldsymbol{x};\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\star}$ and suppose that $p_\mathrm{d}(\boldsymbol{x}) = p_\mathrm{m}(\boldsymbol{x};\boldsymbol{\theta}_\star)$ as implied by the assumptions in Section 4.

It is too difficult to analytically derive the optimal noise distribution in general. To alleviate this problem, we make the following assumptions:

(B1) The noise distribution $p_\mathrm{n}$ is contained in $\mathcal{P}_{\epsilon_0}$ defined as follows: for sufficiently small positive constant $\epsilon_0 \ll 1$,

$$\mathcal{P}_{\epsilon_0} = \Big\{p : \Omega \to [0,\infty) \Big| \int_\Omega p(\boldsymbol{x})\, d\mu(\boldsymbol{x}) = 1, \\ \exists h, \exists \epsilon \in (0,\epsilon_0),\ \frac{p_\mathrm{d}(\boldsymbol{x})}{p(\boldsymbol{x})} = 1 + \epsilon h(\boldsymbol{x}), \|h\|_\infty \le 1\Big\}.$$

(B2) $\varphi$ belongs to $\mathcal{C}^3$.

(B3) $p_\mathrm{n}(\boldsymbol{y})$ is normalized so that $\int p_\mathrm{n}(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = 1$.

(B4) Fisher information matrix $\boldsymbol{I}_\mathrm{F}$ is positive definite.

Assumption (B1) is striking, and enables us to have a first-order expression of MSE with respect to $\epsilon$ whose details are given in Appendix I. The following theorem presents the optimal noise distribution as the minimizer of the first-order expression:

**Theorem 3** (Optimal noise distribution)**.** *Under Assumptions (B1-4), the optimal noise distribution minimizing the first-order approximation of MSE is given by*

$$p_\mathrm{n}^\mathrm{opt}(\boldsymbol{y}) = \frac{p_\mathrm{d}(\boldsymbol{y})\|\boldsymbol{I}_\mathrm{F}^{-1}\boldsymbol{g}_\mathrm{m}(\boldsymbol{y})\|}{Z_\mathrm{n}}, \tag{18}$$

*where* $Z_\mathrm{n} := \mathbb{E}_\mathrm{d}[\|\boldsymbol{I}_\mathrm{F}^{-1}\boldsymbol{g}_\mathrm{m}(\boldsymbol{X})\|]$ *is the normalizing constant.*

The proof is given in Appendix I. The optimal noise distribution $p_\mathrm{n}^\mathrm{opt}$ is exactly the same as the one derived in Chehab et al. [2022] and is clearly different from the data distribution $p_\mathrm{d}$. As depicted in Figure 1, it seems to emphasize the region where $p_\mathrm{d}$ strongly changes. Interestingly, $p_\mathrm{n}^\mathrm{opt}(\boldsymbol{y})$ is optimal for *any* choice of convex functions $\varphi$.

It should be noted that there is a gap in the optimal noise distributions between Proposition 2 and Theorem 3: Theorem 3 shows that the optimal distribution is $p_\mathrm{n}^\mathrm{opt}$, while Proposition 2 indicates that the data distribution $p_\mathrm{d}$ is optimal in the limit of $\nu \to \infty$. In order to bridge $p_\mathrm{d}$ and $p_\mathrm{n}^\mathrm{opt}$, we consider to take the following constraint into account in the optimization problem:

$$\int \frac{(p_\mathrm{d}(\boldsymbol{x}) - p_\mathrm{n}(\boldsymbol{x}))^2}{p_\mathrm{n}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \le C_\epsilon, \tag{19}$$

where $C_\epsilon$ denotes a non-negative constant. The left-hand side on (19) is the Pearson $\chi^2$-divergence, and thus (19) means that $p_\mathrm{n}$ is optimized in a neighborhood of $p_\mathrm{d}$ more explicitly than Theorem 3 as well as Chehab et al. [2022]. The optimal noise distribution is established under this constraint in the following theorem:

**Theorem 4** (Optimal noise distribution in a neighborhood of $p_\mathrm{d}$)**.** *Assume that* $(1+C_\epsilon)/Z_\mathrm{n} \le \int p_\mathrm{d}(\boldsymbol{x})/\|\boldsymbol{I}_\mathrm{F}^{-1}\boldsymbol{g}_\mathrm{m}(\boldsymbol{x})\|\mathrm{d}\boldsymbol{x}$ *holds. Then, the optimal noise distribution minimizing the first-order approximation of MSE under constraint* (19) *is given as follows: For* $0 < \lambda < \infty$*,*

$$\check{p}_\mathrm{n}^\mathrm{opt}(\boldsymbol{y};\lambda) \propto \sqrt{p_\mathrm{n}^\mathrm{opt}(\boldsymbol{y})^2 + \lambda p_\mathrm{d}(\boldsymbol{y})^2}, \tag{20}$$

*where* $\lambda$ *is determined by* $C_\epsilon$ *and satisfies* (47)*. When* $\lambda = \infty$*,* $\check{p}_\mathrm{n}^\mathrm{opt}(\boldsymbol{y};\lambda) = p_\mathrm{d}(\boldsymbol{y})$.

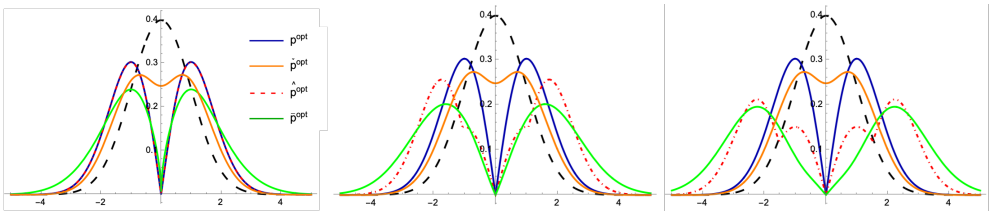
The proof provided in Appendix J is substantially different from that of Chehab et al. [2022], where a novel lemma is established. Interestingly, $\check{p}_\mathrm{n}^\mathrm{opt}(\boldsymbol{y};\lambda)$ is a combination of $p_\mathrm{n}^\mathrm{opt}(\boldsymbol{y})$ and $p_\mathrm{d}(\boldsymbol{y})$. If $\lambda = 0$, then $\check{p}_\mathrm{n}^\mathrm{opt}$ reduces to $p_\mathrm{n}^\mathrm{opt}$, while it approaches $p_\mathrm{d}$ as $\lambda$ increases. This result might give some insight to the gap between the noise distribution in Chehab et al. [2022] and the common practical belief that $p_\mathrm{d}$ is the optimal noise distribution mentioned in Section 1.

Finally, we propose two constraints for robustness and derive the optimal noise distributions under them. Theorem 3 and Figure 1 imply that $p_\mathrm{n}^\mathrm{opt}$ has a wider support than $p_\mathrm{d}$, and thus distribution ratio $r_\mathrm{m}(\boldsymbol{x};\boldsymbol{\alpha}_\star) = p_\mathrm{d}(\boldsymbol{x})/p_\mathrm{n}^\mathrm{opt}(\boldsymbol{x}) \propto \|\boldsymbol{I}_\mathrm{F}^{-1}\boldsymbol{g}_\mathrm{m}(\boldsymbol{x})\|^{-1}$ can approach zero when outliers $\boldsymbol{x}_\mathrm{o}$ go to the tails of $p_\mathrm{d}$. Based on this implication, the condition (12) for B-robustness can be simplified and generalized as follows: For all $\boldsymbol{x}$,

$$\frac{p_\mathrm{d}(\boldsymbol{x})}{p_\mathrm{n}(\boldsymbol{x})}\|\boldsymbol{g}_\mathrm{m}(\boldsymbol{x})\|^\gamma \le C', \tag{21}$$

where $C' > 0$ and $\gamma > 0$ are a positive constant and a tuning parameter, respectively. A larger exponent $\gamma$ increases the divergent speed. Therefore, if (21) holds even for larger $\gamma$, BCE could be more robust. Motivated by this idea, we propose to add (21) as a constraint in optimizing $p_\mathrm{n}$. Furthermore, inspired by (21), we propose another constraint of an integral form, which yields a smoother optimal noise distribution, as follows:

$$\int \frac{p_\mathrm{d}(\boldsymbol{x})}{p_\mathrm{n}(\boldsymbol{x})}\|\boldsymbol{g}_\mathrm{m}(\boldsymbol{x})\|^\gamma \mathrm{d}\boldsymbol{x} \le C'', \tag{22}$$



(a) $\gamma = 1$ (b) $\gamma = 3$ (c) $\gamma = 5$

Figure 1: $p_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y})$, $\check{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y};\lambda)$, $\hat{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y};\gamma)$ and $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y};\lambda,\gamma)$ in mean estimation of the Gaussian distribution (black dotted line) when $\lambda = 0.5$.

where $C''$ is a positive constant. The following theorem 5 establishes the optimal noise distribution under each of the robust constraints:

**Theorem 5** (Optimal noise distribution with robustness)**.** *Under an additional constraint* (21)*, the optimal noise distribution is given*

$$\hat{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y};\gamma) = \max(p_{\mathrm{d}}(\boldsymbol{y})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{y})\|^{\gamma}/C', \lambda p_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y})), \quad (23)$$

*where $\lambda$ is chosen so that $\int \hat{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y};\lambda,\gamma)\mathrm{d}\boldsymbol{y} = 1$.*

*Next, we replace constraint* (21) *by* (22) *and assume that $\int \sqrt{p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}}\mathrm{d}\boldsymbol{x} \leq \sqrt{C''}$. Then, if $0 < \lambda < \infty$, the optimal noise distribution $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}$ is given by*

$$\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{x};\lambda,\gamma) \propto \sqrt{p_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y})^2 + \lambda p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}}, \quad (24)$$

*where $\lambda$ is determined by $C''$ and satisfies* (49)*. When $\lambda = \infty$, $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{x};\gamma) \propto \sqrt{p_{\mathrm{d}}(\boldsymbol{x})}\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma/2}$.*

The proof is given in Appendix K and includes another lemma which has not been seen in Chehab et al. [2022]. Compared with the optimal noise distribution in Theorem 3, $\hat{p}_{\mathrm{n}}^{\mathrm{opt}}$ and $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}$ in (24) have a wider support that $p_{\mathrm{n}}^{\mathrm{opt}}$. This wide support is important in robust estimation because it prevents the ratio $p_{\mathrm{d}}(\boldsymbol{x})/p_{\mathrm{n}}(\boldsymbol{x})$ from diverging when outliers go to the tails of $p_{\mathrm{d}}(\boldsymbol{x})$. Indeed, Figure 1 illustrates $\hat{p}_{\mathrm{n}}^{\mathrm{opt}}$ and $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}$ have a wider support than $p_{\mathrm{n}}^{\mathrm{opt}}$ in Gaussian mean estimation.

# 6 NUMERICAL EXPERIMENTS

This section numerically investigates the robustness of $\tau$-NCE and $\rho$-NCE without model selection, and demonstrates the performance of BCE with the optimal noise distribution derived under a robustness constraint. The experimental results with model selection based on cross-validation are given in Appendix L.

## 6.1 ROBUSTNESS OF BCE

We first generated the contaminated data from

$$(1-\varepsilon)p_{\mathrm{d}}(\boldsymbol{x}) + \varepsilon\bar{\delta}(\boldsymbol{x}), \quad (25)$$

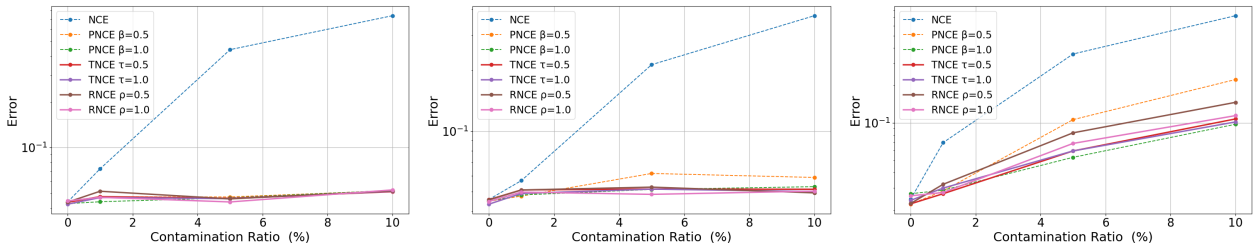
where $\bar{\delta}(\boldsymbol{x})$ denotes a probability distribution of outliers. For $p_{\mathrm{d}}$, we employed three kinds of distributions: Gaussian, log-normal, and Poisson distributions. Distributions in the same family were chosen for $\bar{\delta}(\boldsymbol{x})$ and $p_{\mathrm{n}}(\boldsymbol{x})$. With unnormalized models, we estimated the ten-dimensional mean parameter vector for Gaussian and log-normal distributions, and the single parameter for Poisson distribution. The performance measure is RMSE to the true parameters. The details of the experiments are presented in Appendix L.

The results are presented in Fig. 2 when $\varepsilon = 0, 0.01, 0.05, 0.1$. For all data, both $\tau$-NCE and $\rho$-NCE are robust against outliers, and compare favorably with PNCE. On the other hand, the estimation error of NCE quickly increases as the contamination ratio $\varepsilon$ increases. These results support our theoretical analysis for B-robustness in Section 4.1 and possibly imply that BCE is robust under a variety of convex functions $\varphi$. As implied in Corollary 1, Figure 3 illustrates that the robustness is often improved as $\nu$ increases.

## 6.2 OPTIMAL NOISE DISTRIBUTION

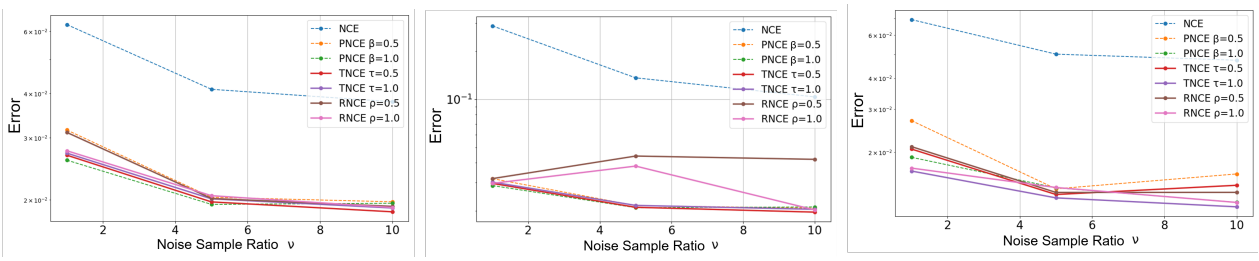
Next, we investigate whether the optimal noise distribution $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{x};\lambda,\gamma)$ in (23) produces robust parameter estimation. To this end, we generated one-dimensional data by the contaminated model (25) where $N(0,1)$ and $N(\mu_{\mathrm{o}}, 0.1^2)$ were employed for $p_{\mathrm{d}}$ and $\bar{\delta}$, respectively. The parameter $\mu_{\mathrm{o}}$ in $N(\mu_{\mathrm{o}}, 0.1^2)$ can be interpreted as the center of outliers. For comparison, $p_{\mathrm{d}}$ and $p_{\mathrm{n}}^{\mathrm{opt}}$ were used as noise distribution $p_{\mathrm{n}}$. Noise data samples were generated using the inverse function method for $p_{\mathrm{n}}^{\mathrm{opt}}$ and slice sampling for $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}$, respectively. As in Section 6.1, we estimated the mean value in the Gaussian distribution $N(0,1)$ by PNCE with $\beta = 1$.

Figure 4 shows estimation errors when the contamination ratio $\varepsilon$ is from $0.01$ to $0.05$. The optimal noise distribution $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{x};\lambda,\gamma)$ produces very robust estimates, while the estimation by $p_{\mathrm{d}}$ and $p_{\mathrm{n}}^{\mathrm{opt}}$ is strongly hampered by the more contamination of outliers. This robustness of $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{x};\lambda,\gamma)$ is more clearly emphasized when $\mu_{\mathrm{o}}$ is located around the heavier-tails of the data distribution $p_{\mathrm{d}}$ (Figure 4(b)). Estimation errors over the noise sample ratio $\hat{\nu}$ are also presented



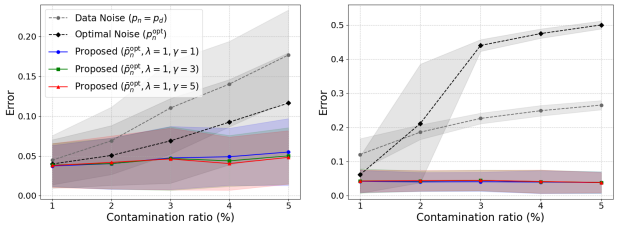

(a) Gaussian data (b) Log-normal data (c) Poisson data

Figure 2: Estimation errors over outlier ratio $\varepsilon$ when $(n, n_{\mathrm{y}}) = (1000, 1000)$. Each point indicates the averaged errors over 100 runs, and the vertical axis is plotted on a logarithm scale.



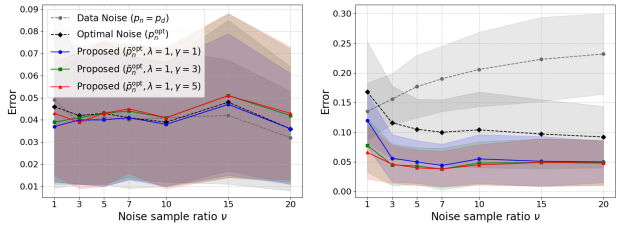

(a) Gaussian data (b) Log-normal data (c) Poisson data

Figure 3: Estimation errors over sample ratio $\hat{\nu} = n_{\mathrm{y}}/n$ when $(n, \varepsilon) = (3000, 0.01)$.




(a) $\mu_{\mathrm{o}} = 3$ (b) $\mu_{\mathrm{o}} = 10$

Figure 4: Estimation errors over contamination ratio $\varepsilon$ when $(n, n_{\mathrm{y}}) = (500, 2500)$. Each point is the average of estimator errors over 100 runs. The shaded regions represent the standard errors.




(a) $\varepsilon = 0$ (no outliers) (b) $\varepsilon = 0.05$

Figure 5: Estimation errors over sample ratio $\hat{\nu}$ when $(\mu_{\mathrm{o}}, n) = (3, 500)$.

in Figure 5.

## 7 CONCLUSION

This paper proposed a general framework for learning unnormalized models based on the Bregman divergence. Our framework includes existing methods as special cases, and enables us to derive novel variants of NCE. We performed asymptotic analysis of our framework in terms of outlier-robustness and efficiency. Our analysis indicates that robust estimation is possible under a variety of convex functions in the Bregman divergence. The proposed framework can achieve the Cramér-Rao bound when the noise distribution is equal to the data one. Inspired by the asymptotic analysis, the optimal noise distribution is also derived under novel constraints for robustness. We numerically demonstrated the robustness of the novel variants of NCE and our optimal noise distribution.


### Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Numbers JP23K28150, JP24K14849 and JP26H02491.

## References


J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society*, 24D(3):179–195, 1975.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3): 200–217, 1967.

C. Ceylan and M. U. Gutmann. Conditional noise-contrastive estimation of unnormalised models. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 726–734. PMLR, 2018.

O. Chehab, A. Gramfort, and A. Hyvärinen. The optimal noise in noise-contrastive learning is not what you think. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 307–316. PMLR, 2022.

H. Chen, G. He, L. Yuan, G. Cui, H. Su, and J. Zhu. Noise contrastive alignment of language models with explicit rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 117784–117812, 2024.

J. Feng, H. Xu, S. Mannor, and S. Yan. Robust logistic regression and classification. *Advances in neural information processing systems (NeurIPS)*, 27, 2014.

R. Gao, E. Nijkamp, D. P. Kingma, Z. Xu, A. M. Dai, and Y. N. Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics (AISTATS)*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

M. U. Gutmann and J. Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 283–290, 2011.

M. U. Gutmann, S. Kleinegesse, and B. Rhodes. Statistical applications of contrastive learning. *Behaviormetrika*, 49 (2):277–301, 2022.

M.U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley, 2009.

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

T. Kanamori and M. Sugiyama. Statistical analysis of distance estimators with density differences and density ratios. *Entropy*, 16(2):921–942, 2014.

P. Lavergne. A Cauchy-Schwarz inequality for expectation of matrices. *Discussion papers, Department of Economics, Simon Fraser University*, 2008.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 1998.

X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 2020.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

T. Matsuda, M. Uehara, and A. Hyvärinen. Information criteria for non-normalized models. *Journal of Machine Learning Research*, 22:1–33, 2021.

M. Minami and S. Eguchi. Adaptive selection for minimum $\beta$-divergence method. In *Proceedings of the fourth international Symposium on Independent Component Analysis and Blind Source Separation*, 2003.

A. Mnih and Y. W. Teh. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

N. Murata, . Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-Boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481, 2004.

H. Sasaki and T. Takenouchi. Outlier-robust parameter estimation for unnormalized statistical models. *Japanese Journal of Statistics and Data Science*, 2024.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

T. Takenouchi and S. Eguchi. Robustifying AdaBoost by adding the naive error rate. *Neural Computation*, 16(4): 767–787, 2004.

T. Takenouchi and T. Kanamori. Statistical inference with unnormalized discrete models and localized homogeneous divergences. *Journal of Machine Learning Research*, 18(1):1804–1829, 2017.

M. Uehara, T. Kanamori, T. Takenouchi, and T. Matsuda. A unified statistically efficient estimation framework for unnormalized models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 809–819. PMLR, 2020.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge university press, 1998.

L. Wasserman. *All of statistics*. Springer, 2004.

# Bregman contrastive estimation as a general framework for learning unnormalized models: Efficiency, robustness and optimal noise (Supplementary Material)

**Yuto Fujii**[1] **Hiroaki Sasaki**[2] **Takafumi Kanamori**[1,3]

[1]Dept. of Math. & Comp. Sci., Institute of Science Tokyo, Tokyo, Japan
[2]Dept. of Math. Info., Meiji Gakuin University, Kanagawa, Japan
[3]RIKEN, Center for Advanced Intelligence Project, Tokyo, Japan

## A NOTATION

Table 2: List of notations.

| | |
|---|---|
| $\boldsymbol{x},\ \boldsymbol{y}$ | Input and noise data |
| $\|\boldsymbol{x}\|$ | Euclidean norm of a vector $\boldsymbol{x}$ |
| $\nabla_{\boldsymbol{\alpha}}$ | Gradient with respect to $\boldsymbol{\alpha}$ |
| $n,\ n_{\mathrm{y}}$ | Numbers of input and noise data samples |
| $p_{\mathrm{d}}(\boldsymbol{x})$ | Data distribution |
| $\bar{p}_{\mathrm{d}}$ | Contaminated distribution of $p_{\mathrm{d}}$ |
| $p_{\mathrm{n}}(\boldsymbol{y})$ | Noise distribution |
| $p_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y}),\quad \check{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y};\lambda),\quad \hat{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y};\gamma),$ $\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y};\lambda,\gamma)$ | Optimal noise distributions |
| $p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta})$ | Statistical model with a parameter vector $\boldsymbol{\theta}$ |
| $p_{\mathrm{m}}^{0}(\boldsymbol{x};\boldsymbol{\alpha})$ | Unnormalized statistical model with a parameter vector $\boldsymbol{\alpha}=(\boldsymbol{\theta},c)$ |
| $\mathcal{D}:=\mathrm{supp}(p_{\mathrm{m}})$ | Support of $p_{\mathrm{m}}$: $\{\boldsymbol{x}\mid p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})\neq 0\}$ |
| $\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})$ | Gradient of $\log p_{\mathrm{m}}^{0}(\boldsymbol{x};\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$: $\nabla_{\boldsymbol{\alpha}}\log p_{\mathrm{m}}^{0}(\boldsymbol{x};\boldsymbol{\alpha})$ |
| $\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})$ | Gradient of $\log p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}=\boldsymbol{\theta}_{\star}$: $\nabla_{\boldsymbol{\theta}}\log p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta})\vert_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\star}}$ |
| $r(\boldsymbol{x})$ | Ratio of $p_{\mathrm{d}}(\boldsymbol{x})$ to $p_{\mathrm{n}}(\boldsymbol{x})$: $p_{\mathrm{d}}(\boldsymbol{x})/p_{\mathrm{n}}(\boldsymbol{x})$ |
| $r_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha})$ | Ratio of $p_{\mathrm{m}}^{0}(\boldsymbol{x};\boldsymbol{\alpha})$ to $p_{\mathrm{n}}(\boldsymbol{x})$: $p_{\mathrm{m}}^{0}(\boldsymbol{x};\boldsymbol{\alpha})/p_{\mathrm{n}}(\boldsymbol{x})$ |
| $f_C(\boldsymbol{x};\boldsymbol{\alpha})\ (C=0,1)$ | Model for the posterior probability |
| $\nu$ | Ratio of class probabilities: $p(C=1)/p(C=0)$ |
| $\hat{\nu}$ | Sample ratio: $n_{\mathrm{y}}/n$ |
| $\varphi,\ \varphi',\ \varphi''$ | Strictly convex function in the Bregman divergence and its derivatives |
| $\boldsymbol{x}_{\mathrm{o}}$ | Outlier |
| $\varepsilon$ | Contamination ratio of outliers |
| $\delta_{\boldsymbol{x}_{\mathrm{o}}}$ | Dirac delta function with the point mass $\boldsymbol{x}_{\mathrm{o}}$ |
| $\mathbb{E}_{\mathrm{d}},\bar{\mathbb{E}}_{\mathrm{d}},\mathbb{E}_{\mathrm{n}}$ | Expectations over $p_{\mathrm{d}}$, $\bar{p}_{\mathrm{d}}$ and $p_{\mathrm{n}}$ |
| $\mathcal{L}_{\varphi}^{\mathrm{SB}}(\boldsymbol{\alpha}),\ \bar{\mathcal{L}}_{\varphi}^{\mathrm{SB}}(\boldsymbol{\alpha})$ | Risk function of BCE and its contaminated risk using $\bar{p}_{\mathrm{d}}$ |
| $\boldsymbol{\alpha}_{\star},\boldsymbol{\alpha}_{\varepsilon}$ | Minimizers of $\mathcal{L}_{\varphi}^{\mathrm{SB}}(\boldsymbol{\alpha})$ and $\bar{\mathcal{L}}_{\varphi}^{\mathrm{SB}}(\boldsymbol{\alpha})$ |
| $\boldsymbol{\Lambda}_{\varphi},\boldsymbol{\Lambda}_{\mathrm{NCE}}$ | Asymptotic variances of BCE and NCE |
| $\boldsymbol{I}_{\mathrm{F}}$ | Fisher information matrix |
| $f_C^{\star}(\boldsymbol{u})$ | Model with the optimal parameter: $f_C(\boldsymbol{x};\boldsymbol{\alpha}_{\star})$ |
| $\mathcal{C}^k$ | Space of continuous and $k$-times differentiable functions |

## B DERIVATION OF (4)

Here, we first express $p(\boldsymbol{u}) = p_0\{p_{\mathrm{d}}(\boldsymbol{u}) + \nu p_{\mathrm{n}}(\boldsymbol{u})\}$ where $p_0 := p(C=0)$. Then, (3) can be written as

$$
\begin{aligned}
&-\int p(\boldsymbol{u}) \sum_{C\in\{0,1\}} [\varphi(f_C(\boldsymbol{u};\boldsymbol{\alpha})) - \varphi'(f_C(\boldsymbol{u};\boldsymbol{\alpha}))(p(C|\boldsymbol{u}) - f_C(\boldsymbol{u};\boldsymbol{\alpha}))]\,\mathrm{d}\boldsymbol{u} \\
&= p_0 \int (p_{\mathrm{d}}(\boldsymbol{u}) + \nu p_{\mathrm{n}}(\boldsymbol{u})) \sum_{C\in\{0,1\}} [-\varphi(f_C(\boldsymbol{u};\boldsymbol{\alpha})) + \varphi'(f_C(\boldsymbol{u};\boldsymbol{\alpha}))f_C(\boldsymbol{u};\boldsymbol{\alpha})]\mathrm{d}\boldsymbol{u} \\
&\quad - \sum_{C\in\{0,1\}} p(C) \int p(\boldsymbol{u}|C)\varphi'(f_C(\boldsymbol{u};\boldsymbol{\alpha}))\mathrm{d}\boldsymbol{u} \\
&= p_0\Big(\mathbb{E}_{\mathrm{d}}\Big[\sum_{C\in\{0,1\}} \{-\varphi(f_C(\boldsymbol{X};\boldsymbol{\alpha})) + \varphi'(f_C(\boldsymbol{X};\boldsymbol{\alpha}))f_C(\boldsymbol{X};\boldsymbol{\alpha})\}\Big] \\
&\quad + \nu\mathbb{E}_{\mathrm{n}}\Big[\sum_{C\in\{0,1\}} \{-\varphi(f_C(\boldsymbol{Y};\boldsymbol{\alpha})) + \varphi'(f_C(\boldsymbol{Y};\boldsymbol{\alpha}))f_C(\boldsymbol{Y};\boldsymbol{\alpha})\}\Big] - \mathbb{E}_{\mathrm{d}}[\varphi'(f_0(\boldsymbol{X};\boldsymbol{\alpha}))] - \nu\mathbb{E}_{\mathrm{n}}[\varphi'(f_1(\boldsymbol{Y};\boldsymbol{\alpha}))]\Big) \\
&= p_0\Big(-\mathbb{E}_{\mathrm{d}}\Big[\varphi(f_0(\boldsymbol{X};\boldsymbol{\alpha})) + \varphi(f_1(\boldsymbol{X};\boldsymbol{\alpha})) - \varphi'(f_0(\boldsymbol{X};\boldsymbol{\alpha}))\{f_0(\boldsymbol{X};\boldsymbol{\alpha}) - 1\} - \varphi'(f_1(\boldsymbol{X};\boldsymbol{\alpha}))f_1(\boldsymbol{X};\boldsymbol{\alpha})\Big] \\
&\quad - \nu\mathbb{E}_{\mathrm{n}}\Big[\varphi(f_0(\boldsymbol{Y};\boldsymbol{\alpha})) + \varphi(f_1(\boldsymbol{Y};\boldsymbol{\alpha})) - \varphi'(f_0(\boldsymbol{Y};\boldsymbol{\alpha}))f_0(\boldsymbol{Y};\boldsymbol{\alpha}) - \varphi'(f_1(\boldsymbol{Y};\boldsymbol{\alpha}))\{f_1(\boldsymbol{Y};\boldsymbol{\alpha}) - 1\}\Big]\Big).
\end{aligned}
$$

Since $f_0(\boldsymbol{u};\boldsymbol{\alpha}) + f_1(\boldsymbol{u};\boldsymbol{\alpha}) = 1$ for all $\boldsymbol{u}$, we finally obtain

$$
\begin{aligned}
\mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) = &-\mathbb{E}_{\mathrm{d}}[\varphi(f_0(\boldsymbol{X};\boldsymbol{\alpha})) + \varphi(f_1(\boldsymbol{X};\boldsymbol{\alpha})) + \{\varphi'(f_0(\boldsymbol{X};\boldsymbol{\alpha})) - \varphi'(f_1(\boldsymbol{X};\boldsymbol{\alpha}))\}f_1(\boldsymbol{X};\boldsymbol{\alpha})] \\
&- \nu\mathbb{E}_{\mathrm{n}}[\varphi(f_0(\boldsymbol{Y};\boldsymbol{\alpha})) + \varphi(f_1(\boldsymbol{Y};\boldsymbol{\alpha})) - \{\varphi'(f_0(\boldsymbol{Y};\boldsymbol{\alpha})) - \varphi'(f_1(\boldsymbol{Y};\boldsymbol{\alpha}))\}f_0(\boldsymbol{Y};\boldsymbol{\alpha})].
\end{aligned}
$$

## C PROOF OF PROPOSITION 1

It follows from the definitions of $\boldsymbol{\alpha}_\star$ and $\boldsymbol{\alpha}_\varepsilon$ that

$$
\nabla_{\boldsymbol{\alpha}}\mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})\big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_\star} = \boldsymbol{0} \quad \text{and} \quad \nabla_{\boldsymbol{\alpha}}\bar{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})\big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_\varepsilon} = \boldsymbol{0}. \tag{26}
$$

A simple calculation shows that $\nabla_{\boldsymbol{\alpha}}\bar{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})$ satisfies

$$
\nabla_{\boldsymbol{\alpha}}\bar{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}}\mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) - \varepsilon\left\{\mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha})f_1(\boldsymbol{X};\boldsymbol{\alpha})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha})] - w_\varphi(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha})f_1(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha})\right\},
$$

where we employed the following formulae:

$$
\nabla_{\boldsymbol{\alpha}} f_0(\boldsymbol{x};\boldsymbol{\alpha}) = f_0(\boldsymbol{x};\boldsymbol{\alpha})f_1(\boldsymbol{x};\boldsymbol{\alpha})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}) \quad \text{and} \quad \nabla_{\boldsymbol{\alpha}} f_1(\boldsymbol{x};\boldsymbol{\alpha}) = -f_0(\boldsymbol{x};\boldsymbol{\alpha})f_1(\boldsymbol{x};\boldsymbol{\alpha})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}). \tag{27}
$$

Since the influence function is defined by a Gâteaux derivative of $\boldsymbol{\alpha}_\varepsilon$, differentiating the second equation in (26) with respect to $\varepsilon$ at $\varepsilon = 0$ yields

$$
\boldsymbol{H}_\varphi \cdot \mathrm{IF}(\boldsymbol{x}_{\mathrm{o}}) - \mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)f_1(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)] + w_\varphi(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha}_\star)f_1(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x}_{\mathrm{o}};\boldsymbol{\alpha}_\star) = \boldsymbol{0},
$$

where $\boldsymbol{H}_\varphi := \nabla_{\boldsymbol{\alpha}}^2\mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})\big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_\star}$. The invertibility assumption of $\boldsymbol{H}_\varphi$ completes the proof.

## D EXAMPLES OF THE WEIGHT FUNCTION $w_\varphi$

This appendix gives examples of the weight function $w_\varphi$ and illustrates them in Figure 6.

- **PNCE**:

$$
w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}) = \left(f_0(\boldsymbol{x};\boldsymbol{\alpha})^{\beta-1} + f_1(\boldsymbol{x};\boldsymbol{\alpha})^{\beta-1}\right) f_0(\boldsymbol{x};\boldsymbol{\alpha})f_1(\boldsymbol{x};\boldsymbol{\alpha}).
$$

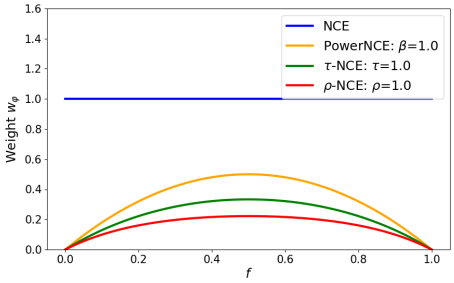



Figure 6: Illustration for the weight function $\{\varphi''(f) + \varphi''(1-f)\}\, f(1-f)$.

- $\tau$-**NCE**:

$$w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}) = \left(\frac{1}{f_0(\boldsymbol{x};\boldsymbol{\alpha}) + \tau} + \frac{1}{f_1(\boldsymbol{x};\boldsymbol{\alpha}) + \tau}\right) f_0(\boldsymbol{x};\boldsymbol{\alpha}) f_1(\boldsymbol{x};\boldsymbol{\alpha}).$$

- $\rho$-**NCE**:

$$w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}) = \left(\frac{1}{(f_0(\boldsymbol{x};\boldsymbol{\alpha}) + \rho)^2} + \frac{1}{(f_1(\boldsymbol{x};\boldsymbol{\alpha}) + \rho)^2}\right) f_0(\boldsymbol{x};\boldsymbol{\alpha}) f_1(\boldsymbol{x};\boldsymbol{\alpha}).$$

## E CONSISTENCY OF BCE

**Proposition 3** (Consistency)**.** *Suppose that Assumptions (A1-6) hold. Then, $\widehat{\boldsymbol{\alpha}}$ converges to $\boldsymbol{\alpha}_\star$ in probability.*

*Proof.* We essentially follow the consistency proof in Van der Vaart [1998]. By definition of $\boldsymbol{\alpha}_\star$ and Assumptions (A3-4), we have $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_\star$. Thus, it suffices that $\widehat{\boldsymbol{\alpha}}$ converges $\boldsymbol{\alpha}_\star$ in probability.

From Assumption (A5), an event $\{\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\| \geq \epsilon\}$ for each $\epsilon > 0$ implies another event $\{\mathcal{L}_\varphi^{\mathrm{SB}}(\widehat{\boldsymbol{\alpha}}) > \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star) + \delta\}$ for some $\delta > 0$. Thus, we have the following inequality of these events:

$$P(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\| \geq \epsilon) \leq P(\mathcal{L}_\varphi^{\mathrm{SB}}(\widehat{\boldsymbol{\alpha}}) - \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star) > \delta).$$

Since $\widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) > \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\widehat{\boldsymbol{\alpha}})$ holds for all $\boldsymbol{\alpha}$ by definition of $\widehat{\boldsymbol{\alpha}}$,

$$\begin{aligned}
\mathcal{L}_\varphi^{\mathrm{SB}}(\widehat{\boldsymbol{\alpha}}) - \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star) &= \mathcal{L}_\varphi^{\mathrm{SB}}(\widehat{\boldsymbol{\alpha}}) - \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star) + \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star) - \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star) \\
&\leq \mathcal{L}_\varphi^{\mathrm{SB}}(\widehat{\boldsymbol{\alpha}}) - \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\widehat{\boldsymbol{\alpha}}) + \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star) - \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star) \\
&\leq 2 \sup_{\boldsymbol{\alpha}} |\mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}) - \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})|.
\end{aligned}$$

Therefore, Assumption (A6) ensures that $\mathcal{L}_\varphi^{\mathrm{SB}}(\widehat{\boldsymbol{\alpha}}) - \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star)$ converges to zero in probability, indicating the convergence of $\widehat{\boldsymbol{\alpha}}$ to $\boldsymbol{\alpha}_\star$. The proof is completed. □

## F PROOF OF THEOREM 1

By definition of $\widehat{\boldsymbol{\alpha}}$, we have

$$\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})\big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}} = -\frac{1}{n}\sum_{i=1}^{n} w_\varphi(\boldsymbol{x}_i;\widehat{\boldsymbol{\alpha}}) f_1(\boldsymbol{x}_i;\widehat{\boldsymbol{\alpha}}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x}_i;\widehat{\boldsymbol{\alpha}}) + \frac{\hat{\nu}}{n_y}\sum_{i=1}^{n_y} w_\varphi(\boldsymbol{y}_i;\widehat{\boldsymbol{\alpha}}) f_0(\boldsymbol{y}_i;\widehat{\boldsymbol{\alpha}}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{y}_i;\widehat{\boldsymbol{\alpha}}) = \mathbf{0}. \tag{28}$$

The Taylor expansion of (28) around $\boldsymbol{\alpha}_\star$ yields

$$\nabla_{\boldsymbol{\alpha}} \mathcal{L}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha})\big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_\star} + \widehat{\boldsymbol{H}}_\varphi(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star) + O\left(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|^2\right) = \mathbf{0}, \tag{29}$$

where

$$
\begin{aligned}
\widehat{\boldsymbol{H}}_{\varphi} &:= \nabla^2_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_\star} \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left[ \{\nabla_{\boldsymbol{\alpha}} w_{\varphi}(\boldsymbol{x}_i;\boldsymbol{\alpha}_\star)\} f_1^\star(\boldsymbol{x}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x}_i;\boldsymbol{\alpha}_\star)^\top + w_{\varphi}(\boldsymbol{x}_i;\boldsymbol{\alpha}_\star)\{\nabla_{\boldsymbol{\alpha}} f_1^\star(\boldsymbol{x}_i)\} \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x}_i;\boldsymbol{\alpha}_\star)^\top \right. \\
&\quad \left. + w_{\varphi}(\boldsymbol{x}_i;\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{x}_i) \nabla_{\boldsymbol{\alpha}} \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x}_i;\boldsymbol{\alpha}_\star) \right] + \frac{\hat{\nu}}{n_y}\sum_{i=1}^{n_y} \left[ \{\nabla_{\boldsymbol{\alpha}} w_{\varphi}(\boldsymbol{y}_i;\boldsymbol{\alpha}_\star)\} f_0^\star(\boldsymbol{y}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{y}_i;\boldsymbol{\alpha}_\star)^\top \right. \\
&\quad \left. + w_{\varphi}(\boldsymbol{y}_i;\boldsymbol{\alpha}_\star)\{\nabla_{\boldsymbol{\alpha}} f_0^\star(\boldsymbol{y}_i)\} \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{y}_i;\boldsymbol{\alpha}_\star)^\top + w_{\varphi}(\boldsymbol{y}_i;\boldsymbol{\alpha}_\star) f_0^\star(\boldsymbol{y}_i) \nabla_{\boldsymbol{\alpha}} \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{y}_i;\boldsymbol{\alpha}_\star) \right],
\end{aligned} \tag{30}
$$

where $f_0^\star(\boldsymbol{x}) := 1 - f_1^\star(\boldsymbol{x}) = \frac{p_{\mathrm{d}}(\boldsymbol{x})}{p_{\mathrm{d}}(\boldsymbol{x}) + \nu p_{\mathrm{n}}(\boldsymbol{x})}$ and $\nabla_{\boldsymbol{\alpha}} f_1^\star(\boldsymbol{x}) := \nabla_{\boldsymbol{\alpha}} f_1(\boldsymbol{x};\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_\star}$. Since $\nu p_{\mathrm{n}} f_0^\star = p_{\mathrm{d}} f_1^\star$, the expectation of $\widehat{\boldsymbol{H}}_{\varphi}$ over $\mathcal{X}$ and $\mathcal{Y}$ is computed as

$$
\begin{aligned}
\mathbb{E}_{\mathcal{X}\times\mathcal{Y}}[\widehat{\boldsymbol{H}}_{\varphi}] &= \int w_{\varphi}(\boldsymbol{x};\boldsymbol{\alpha}_\star) f_0^\star(\boldsymbol{x}) f_1^\star(\boldsymbol{x}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)^\top p_{\mathrm{d}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\quad + \int w_{\varphi}(\boldsymbol{y};\boldsymbol{\alpha}_\star) f_0^\star(\boldsymbol{y}) f_1^\star(\boldsymbol{y}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{y};\boldsymbol{\alpha}_\star) \boldsymbol{g}_{\mathrm{m}}^\top(\boldsymbol{y};\boldsymbol{\alpha}_\star) \hat{\nu} p_{\mathrm{n}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \\
&= \int \left\{ f_0^\star(\boldsymbol{x}) + \frac{\hat{\nu}}{\nu} f_1^\star(\boldsymbol{x}) \right\} w_{\varphi}(\boldsymbol{x};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{x}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star)^\top p_{\mathrm{d}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\mathrm{d}} \left[ \left\{ f_0^\star(\boldsymbol{X}) + \frac{\hat{\nu}}{\nu} f_1^\star(\boldsymbol{X}) \right\} w_{\varphi}(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top \right].
\end{aligned}
$$

Since $\hat{\nu}$ is assumed to converge to $\nu$, $f_0^\star(\boldsymbol{x}) + \frac{\hat{\nu}}{\nu} f_1^\star(\boldsymbol{x}) \to 1$ as $n \to \infty$ and $n_{\mathrm{y}} \to \infty$. Therefore, the law of large numbers guarantees that $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha}_\star)$ converges to $\boldsymbol{J}_{\varphi} = \mathbb{E}_{\mathrm{d}} \left[ w_{\varphi}(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top \right]$ in probability.

On the other hand, by definition of $\boldsymbol{\alpha}_\star$, the expectation of $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha})$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_\star$ satisfies

$$
\mathbb{E}_{\mathcal{X}\times\mathcal{Y}}[\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha}_\star)] = \boldsymbol{0}, \tag{31}
$$

The variance of $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha})$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_\star$ can be computed as follows:

$$
\begin{aligned}
&\mathrm{Var}_{\mathcal{X}\times\mathcal{Y}}[\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha}_\star)] \\
&= \mathbb{E}_{\mathcal{X}\times\mathcal{Y}} \left[ \left( -\frac{1}{n}\sum_{i=1}^{n} w_{\varphi}(\boldsymbol{X}_i;\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X}_i;\boldsymbol{\alpha}_\star) + \frac{\hat{\nu}}{n_y}\sum_{i=1}^{n_y} w_{\varphi}(\boldsymbol{Y}_i;\boldsymbol{\alpha}_\star) f_0^\star(\boldsymbol{Y}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{Y}_i;\boldsymbol{\alpha}_\star) \right) \right. \\
&\quad \left. \left( -\frac{1}{n}\sum_{i=1}^{n} w_{\varphi}(\boldsymbol{X}_i;\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X}_i;\boldsymbol{\alpha}_\star) + \frac{\hat{\nu}}{n_y}\sum_{i=1}^{n_y} w_{\varphi}(\boldsymbol{Y}_i;\boldsymbol{\alpha}_\star) f_0^\star(\boldsymbol{Y}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{Y}_i;\boldsymbol{\alpha}_\star) \right)^\top \right] \\
&= \frac{1}{n^2} \mathbb{E}_{\mathcal{X}} \left[ \left( \sum_{i=1}^{n} w_{\varphi}(\boldsymbol{X}_i;\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X}_i;\boldsymbol{\alpha}_\star) \right) \left( \sum_{i=1}^{n} w_{\varphi}(\boldsymbol{X}_i;\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X}_i;\boldsymbol{\alpha}_\star) \right)^\top \right] \\
&\quad + \frac{\hat{\nu}^2}{n_y^2} \mathbb{E}_{\mathcal{Y}} \left[ \left( \sum_{i=1}^{n_y} w_{\varphi}(\boldsymbol{Y}_i;\boldsymbol{\alpha}_\star) f_0^\star(\boldsymbol{Y}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{Y}_i;\boldsymbol{\alpha}_\star) \right) \left( \sum_{i=1}^{n_y} w_{\varphi}(\boldsymbol{Y}_i;\boldsymbol{\alpha}_\star) f_0^\star(\boldsymbol{Y}_i) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{Y}_i;\boldsymbol{\alpha}_\star) \right)^\top \right] \\
&\quad - 2\hat{\nu} \mathbb{E}_{\mathrm{d}}[w_{\varphi}(\boldsymbol{X}) f_1^\star(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)] \mathbb{E}_{\mathrm{n}}[w_{\varphi}(\boldsymbol{Y};\boldsymbol{\alpha}_\star) f_0^\star(\boldsymbol{Y}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{Y};\boldsymbol{\alpha}_\star)]^\top \\
&= \frac{1}{n} \mathbb{E}_{\mathrm{d}}[w_{\varphi}(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top] \\
&\quad - \frac{1}{n} \Big( 1 + \frac{1}{\hat{\nu}} \Big) \mathbb{E}_{\mathrm{d}}[w_{\varphi}(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)] \mathbb{E}_{\mathrm{d}}[w_{\varphi}(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]^\top
\end{aligned}
$$

Hence, it follows from the central limit theorem that $\sqrt{n} \nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha}_\star)$ converges to a normal distribution as follows:

$$
\sqrt{n} \nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}^{\mathrm{SB}}_{\varphi}(\boldsymbol{\alpha}_\star) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_{\varphi}), \tag{32}
$$

where

$$\boldsymbol{I}_\varphi := \mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top] - \Big(1+\frac{1}{\nu}\Big)\mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]\mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1(\boldsymbol{X}) \boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]^\top.$$

By the dominant convergence theorem and the consistency of $\widehat{\boldsymbol{\alpha}}$, higher-order terms in(29) can be ignored, which yields

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star) = \sqrt{n}\boldsymbol{H}_\varphi^{-1} \nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_\varphi^{\mathrm{SB}}(\boldsymbol{\alpha}_\star).$$

Thus, Slutsky's theorem [Van der Vaart, 1998] and (32) ensures that $\sqrt{n}\big(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\big)$ converges to a normal distribution as follows:

$$\sqrt{n}\big(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\big) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{J}_\varphi^{-1} \boldsymbol{I}_\varphi \boldsymbol{J}_\varphi^{-1}).$$

This completes the proof.

## G PROOF OF THEOREM 2

As derived in Gutmann and Hyvärinen [2012], the asymptotic variance of NCE is given by

$$\boldsymbol{\Lambda}_{\mathrm{NCE}} := \boldsymbol{I}^{-1}\left(\boldsymbol{I} - \left(1+\frac{1}{\nu}\right)\boldsymbol{m}\boldsymbol{m}^\top\right)\boldsymbol{I}^{-1}, \tag{33}$$

where

$$\boldsymbol{I} := \mathbb{E}_{\mathrm{d}}\left[f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right] \quad \text{and} \quad \boldsymbol{m} := \mathbb{E}_{\mathrm{d}}\left[f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\right].$$

By definition of $\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha})$, we have

$$\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} \log p_{\mathrm{m}}^0(\boldsymbol{u};\boldsymbol{\alpha}) = (\nabla_{\boldsymbol{\theta}} \log p_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\theta})^\top, -1)^\top. \tag{34}$$

Then, by letting a canonical vector be $\boldsymbol{e}_1 := (0,...,0,1)$, it holds $\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}_\star)^\top \boldsymbol{e}_1 = -1$, based on which the second term on the right-hand side of $\boldsymbol{I}_\varphi$ in (14) can be computed as

$$\begin{aligned}
&\mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X}) w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]\mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X}) w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]^\top \\
&= \mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X}) w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top \boldsymbol{e}_1]\mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X}) w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top \boldsymbol{e}_1]^\top \\
&= \mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X}) w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top]\boldsymbol{e}_1\boldsymbol{e}_1^\top \mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X}) w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top] \\
&= \boldsymbol{J}_\varphi \boldsymbol{e}_1 \boldsymbol{e}_1^\top \boldsymbol{J}_\varphi.
\end{aligned} \tag{35}$$

Eq.(35) enables us to express $\boldsymbol{\Lambda}_\varphi$ as

$$\boldsymbol{\Lambda}_\varphi = \boldsymbol{J}_\varphi^{-1}\boldsymbol{I}_\varphi\boldsymbol{J}_\varphi^{-1} = \boldsymbol{J}_\varphi^{-1}\left(\mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{x}) w_\varphi^2(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top]\right)\boldsymbol{J}_\varphi^{-1} - \left(1+\frac{1}{\nu}\right)\boldsymbol{e}_1\boldsymbol{e}_1^\top. \tag{36}$$

Applying a matrix version of the Cauchy-Schwarz inequality [Lavergne, 2008] to the first term in (36) gives

$$\boldsymbol{J}_\varphi^{-1}\left(\mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X}) w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)^2\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top]\right)\boldsymbol{J}_\varphi^{-1} \succeq \mathbb{E}_{\mathrm{d}}\left[f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right]^{-1} = \boldsymbol{I}^{-1}. \tag{37}$$

On the other hand, the second term in (36) can be computed as

$$\boldsymbol{e}_1\boldsymbol{e}_1^\top = \boldsymbol{I}^{-1}\{\boldsymbol{I}(\boldsymbol{e}_1\boldsymbol{e}_1^\top)\boldsymbol{I}\}\boldsymbol{I}^{-1} = \boldsymbol{I}^{-1}(\boldsymbol{m}\boldsymbol{m}^\top)\boldsymbol{I}^{-1},$$

where we applied the following equation derived similarly as (35):

$$\boldsymbol{m}\boldsymbol{m}^\top = \mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]\mathbb{E}_{\mathrm{d}}[f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]^\top = \boldsymbol{I}\boldsymbol{e}_1\boldsymbol{e}_1^\top\boldsymbol{I}. \tag{38}$$

Finally, by combining (36) with (37) and (38), we obtain

$$\boldsymbol{\Lambda}_\varphi \succeq \boldsymbol{I}^{-1} - \left(1+\frac{1}{\nu}\right)\boldsymbol{I}^{-1}(\boldsymbol{m}\boldsymbol{m}^\top)\boldsymbol{I}^{-1} = \boldsymbol{I}^{-1}\left\{\boldsymbol{I} - \left(1+\frac{1}{\nu}\right)\boldsymbol{m}\boldsymbol{m}^\top\right\}\boldsymbol{I}^{-1} = \boldsymbol{\Lambda}_{\mathrm{NCE}}.$$

Equality holds if the following equation holds almost everywhere: For all $\boldsymbol{u}$, there exist constant vectors $\boldsymbol{a}, \boldsymbol{b}$ such that

$$w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star)\sqrt{f_1^\star(\boldsymbol{x})}\boldsymbol{a}^\top\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star) + \sqrt{f_1^\star(\boldsymbol{x})}\boldsymbol{b}^\top\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\alpha}_\star) = 0,$$

implying that $w_\varphi(\boldsymbol{u};\boldsymbol{\alpha}_\star)$ is constant by assuming that $\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}_\star)\|$ and $\sqrt{f_1^\star(\boldsymbol{u})}$ are positive.

## H PROOF OF PROPOSITION 2

It follows from the definition of $\boldsymbol{\alpha}_\star$ and Assumptions (A3-4) that $p_{\mathrm{d}}(\cdot) = p^0_{\mathrm{m}}(\cdot;\boldsymbol{\alpha}_\star)$ holds, giving

$$f^\star_0(\boldsymbol{u}) = \frac{p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}_\star)}{p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}_\star)+\nu p_{\mathrm{n}}(\boldsymbol{u})} = \frac{p_{\mathrm{d}}(\boldsymbol{u})}{p_{\mathrm{d}}(\boldsymbol{u})+\nu p_{\mathrm{n}}(\boldsymbol{u})} = \frac{p_{\mathrm{d}}(\boldsymbol{u})}{\nu p_{\mathrm{n}}(\boldsymbol{u})} + O(\nu^{-2}) \tag{39}$$

$$f^\star_1(\boldsymbol{u}) = \frac{\nu p_{\mathrm{n}}(\boldsymbol{u})}{p^0_{\mathrm{m}}(\boldsymbol{u};\boldsymbol{\alpha}_\star)+\nu p_{\mathrm{n}}(\boldsymbol{u})} = \frac{\nu p_{\mathrm{n}}(\boldsymbol{u})}{p_{\mathrm{d}}(\boldsymbol{u})+\nu p_{\mathrm{n}}(\boldsymbol{u})} = 1 - \frac{p_{\mathrm{d}}(\boldsymbol{u})}{\nu p_{\mathrm{n}}(\boldsymbol{u})} + O(\nu^{-2}), \tag{40}$$

By assumption that $\varphi$ is of class $\mathcal{C}^2$, $\varphi''$ is bounded on $[0,1]$. Thus, we can define the limit values as $C_0 = \lim_{z\to 0}\varphi''(z)$ and $C_1 = \lim_{z\to 1}\varphi''(z)$. Then, the weight function $w_\varphi$ has the following order expression:

$$w_\varphi(\boldsymbol{u};\boldsymbol{\alpha}_\star) = \frac{(C_0+C_1)p_{\mathrm{d}}(\boldsymbol{u})}{\nu p_{\mathrm{n}}(\boldsymbol{u})} + O(\nu^{-2}). \tag{41}$$

Substituting (40) and (41) into $\boldsymbol{J}_\varphi$ gives

$$\boldsymbol{J}_\varphi = (C_0+C_1)\mathbb{E}_{\mathrm{d}}\left[\frac{p_{\mathrm{d}}(\boldsymbol{X})}{\nu p_{\mathrm{n}}(\boldsymbol{X})}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right] + O(\nu^{-2}) = \frac{C_0+C_1}{\nu}\tilde{\boldsymbol{J}} + O(\nu^{-2}). \tag{42}$$

Similarly, regarding the first term in $\boldsymbol{I}_\varphi$, we have

$$\mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)^2 f^\star_1(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top] = \frac{(C_0+C_1)^2}{\nu^2}\mathbb{E}_{\mathrm{d}}\left[\left(\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})}\right)^2\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)^\top\right] + O(\nu^{-3}),$$

while the second term is written as

$$\begin{aligned}&\left(1+\frac{1}{\nu}\right)\mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)f^\star_1(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]\mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)f^\star_1(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)]\\ &= \frac{(C_0+C_1)^2}{\nu^2}\mathbb{E}_{\mathrm{d}}\left[\left(\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})}\right)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\right]\mathbb{E}_{\mathrm{d}}\left[\left(\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})}\right)\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{\alpha}_\star)\right]^\top + O(\nu^{-3}).\end{aligned}$$

By using these expressions in $\boldsymbol{I}_\varphi$, we obtain

$$\boldsymbol{I}_\varphi = \frac{(C_0+C_1)^2}{\nu^2}\tilde{\boldsymbol{I}} + O(\nu^{-3}). \tag{43}$$

Since

$$\boldsymbol{J}_\varphi^{-1} = \left\{\frac{C_0+C_1}{\nu}\right\}^{-1}\tilde{\boldsymbol{J}}^{-1} + O(\nu^{-2}),$$

we have $\boldsymbol{J}_\varphi^{-1}\boldsymbol{I}_\varphi\boldsymbol{J}_\varphi^{-1} = \tilde{\boldsymbol{J}}^{-1}\tilde{\boldsymbol{I}}\tilde{\boldsymbol{J}}^{-1} + O(\nu^{-2})$, which completes the proof.

## I PROOF OF THEOREM 3

Since $p_{\mathrm{d}}(\boldsymbol{x}) = p_{\mathrm{m}}(\boldsymbol{x};\boldsymbol{\theta}_\star)$, $f^\star_1(\boldsymbol{x})$ can be represented as in (13). Then, $\frac{p_{\mathrm{d}}(\boldsymbol{x})}{p_{\mathrm{n}}(\boldsymbol{x})} = 1+\epsilon h(\boldsymbol{x})$ in Assumption (B1) gives the following order expression:

$$f^\star_1(\boldsymbol{x}) = \frac{\nu p_{\mathrm{n}}(\boldsymbol{x})}{p_{\mathrm{d}}(\boldsymbol{x})+\nu p_{\mathrm{n}}(\boldsymbol{x})} = \frac{1}{1+\frac{1}{\nu}+\frac{\epsilon h(\boldsymbol{x})}{\nu}} = \frac{\nu}{\nu+1} - \frac{\nu}{(\nu+1)^2}\epsilon h(\boldsymbol{x}) + O(\epsilon^2)$$

$$f^\star_0(\boldsymbol{x}) = \frac{p_{\mathrm{d}}(\boldsymbol{x})}{p_{\mathrm{d}}(\boldsymbol{x})+\nu p_{\mathrm{n}}(\boldsymbol{x})} = \frac{\frac{1}{\nu}+\epsilon\frac{h(\boldsymbol{x})}{\nu}}{1+\frac{1}{\nu}+\frac{\epsilon h(\boldsymbol{x})}{\nu}} = \frac{1}{\nu+1} + \frac{\nu}{(\nu+1)^2}\epsilon h(\boldsymbol{x}) + O(\epsilon^2).$$

These order expressions enable us to represent $w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star)$ as

$$\begin{aligned}
w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star) &= \{\varphi''(f_0^\star(\boldsymbol{x})) + \varphi''(f_1^\star(\boldsymbol{x}))\} f_0^\star(\boldsymbol{x}) f_1^\star(\boldsymbol{x}) \\
&= \left\{\varphi''\left(\frac{1}{\nu+1}\right) + \varphi'''\left(\frac{1}{\nu+1}\right)\frac{\nu}{(\nu+1)^2}\epsilon h(\boldsymbol{x}) + \varphi''\left(\frac{\nu}{\nu+1}\right) - \varphi'''\left(\frac{\nu}{\nu+1}\right)\frac{\nu}{(\nu+1)^2}\epsilon h(\boldsymbol{x}) + O(\epsilon^2)\right\} \\
&\quad \times \left\{\frac{\nu}{(\nu+1)^2} + \frac{\nu^2-\nu}{(\nu+1)^3}\epsilon h(\boldsymbol{x}) + O(\epsilon^2)\right\} \\
&= C_1 + C_2 \epsilon h(\boldsymbol{x}) + O(\epsilon^2),
\end{aligned}$$

where

$$\begin{aligned}
C_1 &:= \left\{\varphi''\left(\frac{\nu}{\nu+1}\right) + \varphi''\left(\frac{1}{\nu+1}\right)\right\}\frac{\nu}{(\nu+1)^2} \\
C_2 &:= \left\{\varphi''\left(\frac{\nu}{\nu+1}\right) + \varphi''\left(\frac{1}{\nu+1}\right)\right\}\frac{\nu^2-\nu}{(\nu+1)^3} + \left\{\varphi'''\left(\frac{1}{\nu+1}\right) - \varphi'''\left(\frac{\nu}{\nu+1}\right)\right\}\frac{\nu^2}{(\nu+1)^4}.
\end{aligned}$$

In order to compute $\boldsymbol{I}_\varphi$ and $\boldsymbol{J}_\varphi$, we substitute this expression of $w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star)$ into

$$\begin{aligned}
w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{x}) &= \left\{C_1 + C_2\epsilon h(\boldsymbol{x}) + O(\epsilon^2)\right\}\left\{\frac{\nu}{\nu+1} - \frac{\nu}{(\nu+1)^2}\epsilon h(\boldsymbol{x}) + O(\epsilon^2)\right\} \\
&= \frac{C_1\nu}{\nu+1} + \left\{\frac{C_2\nu}{\nu+1} - \frac{C_1\nu}{(\nu+1)^2}\right\}\epsilon h(\boldsymbol{x}) + O(\epsilon^2) \\
w_\varphi(\boldsymbol{x};\boldsymbol{\alpha}_\star)^2 f_1^\star(\boldsymbol{x}) &= \left\{C_1^2 + 2C_1C_2\epsilon h(\boldsymbol{x}) + O(\epsilon^2)\right\}\left\{\frac{\nu}{\nu+1} - \frac{\nu}{(\nu+1)^2}\epsilon h(\boldsymbol{x}) + O(\epsilon^2)\right\} \\
&= \frac{C_1^2\nu}{\nu+1} + \left\{\frac{2C_1C_2\nu}{\nu+1} - \frac{C_1^2\nu}{(\nu+1)^2}\right\}\epsilon h(\boldsymbol{x}) + O(\epsilon^2).
\end{aligned}$$

Then, $\boldsymbol{I}_\varphi$ and $\boldsymbol{J}_\varphi$ are expressed as

$$\begin{aligned}
\boldsymbol{I}_\varphi &= \mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star)^2 f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})^\top] \\
&\quad - \mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})]\mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})]^\top \\
&= C_1^2\frac{\nu}{\nu+1}\boldsymbol{I}_{\mathrm{F}} + \left\{\frac{2C_1C_2\nu}{\nu+1} - \frac{C_1^2\nu}{(\nu+1)^2}\right\}\boldsymbol{A}(\epsilon) + O(\epsilon^2) \\
\boldsymbol{J}_\varphi &= \mathbb{E}_{\mathrm{d}}[w_\varphi(\boldsymbol{X};\boldsymbol{\alpha}_\star) f_1^\star(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})^\top] = C_1\frac{\nu}{\nu+1}\boldsymbol{I}_{\mathrm{F}} + \left\{\frac{C_2\nu}{\nu+1} - \frac{C_1\nu}{(\nu+1)^2}\right\}\boldsymbol{A}(\epsilon) + O(\epsilon^2),
\end{aligned}$$

where $\boldsymbol{A}(\epsilon) := \epsilon\mathbb{E}_{\mathrm{d}}[h(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})^\top]$ and we applied $\mathbb{E}_{\mathrm{d}}[\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})] = 0$. Since

$$\begin{aligned}
\boldsymbol{J}_\varphi^{-1} &= \left\{C_1\frac{\nu}{\nu+1}\boldsymbol{I}_{\mathrm{F}} + \left\{\frac{C_2\nu}{\nu+1} - \frac{C_1\nu}{(\nu+1)^2}\right\}\boldsymbol{A}(\epsilon) + O(\epsilon^2)\right\}^{-1} \\
&= \frac{\nu+1}{C_1\nu}\boldsymbol{I}_{\mathrm{F}}^{-1} - \left(\frac{C_2(\nu+1)}{C_1^2\nu} - \frac{1}{C_1\nu}\right)\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{A}(\epsilon)\boldsymbol{I}_{\mathrm{F}}^{-1} + O(\epsilon^2),
\end{aligned}$$

we compute the asymptotic variance $\boldsymbol{\Lambda}_\varphi$ as follows:

$$\boldsymbol{\Lambda}_\varphi = \boldsymbol{J}_\varphi^{-1}\boldsymbol{I}_\varphi\boldsymbol{J}_\varphi^{-1} = \frac{\nu+1}{\nu}\boldsymbol{I}_{\mathrm{F}}^{-1} + \frac{1}{\nu}\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{A}(\epsilon)\boldsymbol{I}_{\mathrm{F}}^{-1} + O(\epsilon^2).$$

The first-order approximation of $\mathrm{MSE} = \mathrm{Tr}(\boldsymbol{\Lambda}_\varphi)/n$ is given by

$$\begin{aligned}
\mathrm{MSE} &= \frac{\nu+1}{n\nu}\mathrm{Tr}(\boldsymbol{I}_{\mathrm{F}}^{-1}) + \frac{1}{n\nu}\mathrm{Tr}(\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{A}(\epsilon)\boldsymbol{I}_{\mathrm{F}}^{-1}) + O(\epsilon^2) && (44) \\
&= \frac{\nu+1}{n\nu}\mathrm{Tr}(\boldsymbol{I}_{\mathrm{F}}^{-1}) + \frac{\epsilon}{n\nu}\mathbb{E}_{\mathrm{d}}\left[\left\{\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})} - 1\right\}\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})\|^2\right], && (45)
\end{aligned}$$

where since $\boldsymbol{I}_{\mathrm{F}}$ is symmetric and $\epsilon h(\boldsymbol{x}) = \frac{p_{\mathrm{d}}(\boldsymbol{x})}{p_{\mathrm{n}}(\boldsymbol{x})} - 1$, we computed the second term as

$$\mathrm{Tr}(\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{A}(\epsilon)\boldsymbol{I}_{\mathrm{F}}^{-1}) = \epsilon\mathbb{E}_{\mathrm{d}}\left[h(\boldsymbol{X})\mathrm{Tr}\left(\{\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})\}\{\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})\boldsymbol{I}_{\mathrm{F}}^{-1}\}^{\top}\right)\right] = \epsilon\mathbb{E}_{\mathrm{d}}\left[\left\{\frac{p_{\mathrm{d}}(\boldsymbol{X})}{p_{\mathrm{n}}(\boldsymbol{X})} - 1\right\}\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})\|^2\right].$$

Next, we derive the optimal noise distribution as the minimizer of the second term on the right-hand side of (45) with a constraint $\int p_{\mathrm{n}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. Hence, we optimize the following Lagrangian:

$$\min_{p_{\mathrm{n}}}\left[\int \frac{p_{\mathrm{d}}(\boldsymbol{x})^2}{p_{\mathrm{n}}(\boldsymbol{x})}\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2\mathrm{d}\boldsymbol{x} + \lambda\int p_{\mathrm{n}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right], \tag{46}$$

where $\lambda$ denotes the Lagrange multiplier. By computing the variational derivative and setting it to zero, we have

$$p_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{y}) \propto p_{\mathrm{d}}(\boldsymbol{y})\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{y})\|,$$

which completes the proof.

## J PROOF OF THEOREM 4

Our proof relies on the following Lemma whose proof is provided below:

**Lemma 1.** *Let $q(\boldsymbol{x})$ be a probability density and let $u(\boldsymbol{x}) \geq 0$ be an integrable function. Consider the optimization problem*

$$\min_{p:\mathrm{p.d.f}}\int \frac{q(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \quad \textit{such that} \quad \int \frac{u(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \leq C,$$

*where $C$ is a positive constant. Then,*

*(a) If $C < \left(\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^2$, then the problem is infeasible.*

*(b) If $\left(\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^2 \leq C < \int \frac{u(\boldsymbol{x})^2}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$, then there exists $\lambda_0 \in (0,\infty]$ such that*

$$\left(\int \frac{u(\boldsymbol{x})^2}{\sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}}\mathrm{d}\boldsymbol{x}\right)\left(\int \sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}\mathrm{d}\boldsymbol{x}\right) = C,$$

*and the optimal solution is*

$$p_{\mathrm{opt}}(\boldsymbol{x}) = \frac{\sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}}{\int \sqrt{q(\boldsymbol{z})^2 + \lambda_0 u(\boldsymbol{z})^2}\mathrm{d}\boldsymbol{z}}.$$

*With the convention $\lambda_0 = \infty$, which corresponds to $C = (\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x})^2$, we obtain*

$$p_{\mathrm{opt}}(\boldsymbol{x}) = \frac{u(\boldsymbol{x})}{\int u(\boldsymbol{z})\mathrm{d}\boldsymbol{z}}$$

*(c) If $C \geq \int \frac{u(\boldsymbol{x})^2}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$, then $p_{\mathrm{opt}}(\boldsymbol{x}) = q(\boldsymbol{x})$.*

We first rewrite the Pearson $\chi^2$ divergence in (19) as

$$\int \frac{(p_{\mathrm{d}}(\boldsymbol{x}) - p_{\mathrm{n}}(\boldsymbol{x}))^2}{p_{\mathrm{n}}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \int \frac{p_{\mathrm{d}}(\boldsymbol{x})^2}{p_{\mathrm{n}}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - 1.$$

Then, our optimization problem is formulated as

$$\min_{p_{\mathrm{n}}:\,\mathrm{p.d.f.}}\int \frac{p_{\mathrm{d}}(\boldsymbol{x})^2}{p_{\mathrm{n}}(\boldsymbol{x})}\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2\mathrm{d}\boldsymbol{x} \quad \text{such that} \quad \int \frac{p_{\mathrm{d}}(\boldsymbol{x})^2}{p_{\mathrm{n}}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \leq 1 + C_\epsilon.$$

To apply Lemma 1, we next substitute $q(\boldsymbol{x}) = p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|/Z_{\mathrm{n}}$ and $u(\boldsymbol{x}) = p_{\mathrm{d}}(\boldsymbol{x})$. Since $C_\epsilon$ is a non-negative constant, it holds $(\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x})^2 = 1 \leq 1 + C_\epsilon$ and thus, Case (a) in Lemma 1 is excluded. On the other hand, the left-hand side of the condition in Case (c) of Lemma 1 is given by

$$\int \frac{u(\boldsymbol{x})^2}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = Z_{\mathrm{n}} \int \frac{p_{\mathrm{d}}(\boldsymbol{x})}{\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|}\mathrm{d}\boldsymbol{x}.$$

The assumption excludes Case (c). Hence, the optimal noise distribution is Case (b) and given by

$$\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{x};\gamma,\lambda) \propto \begin{cases} \sqrt{p_{\mathrm{d}}(\boldsymbol{x})^2\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2/Z_{\mathrm{n}}^2 + \lambda p_{\mathrm{d}}(\boldsymbol{x})^2} & 0 < \lambda < \infty \\ \sqrt{p_{\mathrm{d}}(\boldsymbol{x})}\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma/2} & \lambda = \infty, \end{cases}$$

where $\lambda$ is chosen such that

$$1 + C_\epsilon = \left(\int \frac{p_{\mathrm{d}}(\boldsymbol{x})^2}{\sqrt{p_{\mathrm{d}}(\boldsymbol{x})^2\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2/Z_{\mathrm{n}}^2 + \lambda p_{\mathrm{d}}(\boldsymbol{x})^2}}\mathrm{d}\boldsymbol{x}\right)\left(\int \sqrt{p_{\mathrm{d}}(\boldsymbol{x})^2\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2/Z_{\mathrm{n}}^2 + \lambda p_{\mathrm{d}}(\boldsymbol{x})^2}\mathrm{d}\boldsymbol{x}\right). \quad (47)$$

*Proof of Lemma 1.* (a) For any probability density $p$, by Cauchy-Schwarz inequality,

$$\left(\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^2 = \left(\int \frac{u(\boldsymbol{x})}{\sqrt{p(\boldsymbol{x})}}\sqrt{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right)^2 \leq \left(\int \frac{u(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right)\left(\int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) = \int \frac{u(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}.$$

Hence $\int u(\boldsymbol{x})^2/p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \geq (\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x})^2$, so the problem is infeasible when $C < (\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x})^2$.

(c) For any probability density $p$,

$$\int \frac{q(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \int \frac{(q(\boldsymbol{x}) - p(\boldsymbol{x}))^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} + 1 \geq 1,$$

with equality if and only if $p = q$ a.e. If $C \geq \int u(\boldsymbol{x})^2/q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, then $p = q$ is feasible and thus optimal.

(b) Define, for $\lambda > 0$,

$$h(\lambda) := \left(\int \frac{u(\boldsymbol{x})^2}{\sqrt{q(\boldsymbol{x})^2 + \lambda u(\boldsymbol{x})^2}}\mathrm{d}\boldsymbol{x}\right)\left(\int \sqrt{q(\boldsymbol{x})^2 + \lambda u(\boldsymbol{x})^2}\mathrm{d}\boldsymbol{x}\right).$$

The dominated convergence theorem guarantees that $h(\lambda)$ for $0 < \lambda < \infty$ is continuous. Indeed, the inequalities

$$\frac{u(\boldsymbol{x})^2}{\sqrt{q(\boldsymbol{x})^2 + \lambda u(\boldsymbol{x})^2}} \leq \frac{u(\boldsymbol{x})}{\sqrt{\lambda}},$$
$$\sqrt{q(\boldsymbol{x})^2 + \lambda u(\boldsymbol{x})^2} \leq q(\boldsymbol{x}) + \sqrt{\lambda}u(\boldsymbol{x})$$

and the integrability of $q$ and $u$ lead to the continuity of $h(\lambda)$. Since $h(0) = \int u(\boldsymbol{x})^2/q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and $\lim_{\lambda\to\infty} h(\lambda) = (\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x})^2$, there exists $\lambda_0 \in (0,\infty]$ such that $h(\lambda_0) = C$ under the assumption of (b).

Assume first that $\lambda_0 < \infty$. Let $p$ be any feasible probability density. Then,

$$\int \frac{q(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \int \frac{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - \lambda_0 \int \frac{u(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \geq \int \frac{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - \lambda_0 C.$$

By Cauchy-Schwarz inequality,

$$\int \frac{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \int \frac{\left(\sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}\right)^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \geq \frac{\left(\int \sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}\mathrm{d}\boldsymbol{x}\right)^2}{\int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}} = \left(\int \sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}\mathrm{d}\boldsymbol{x}\right)^2.$$

Therefore,

$$\int \frac{q(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \geq \left(\int \sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}\mathrm{d}\boldsymbol{x}\right)^2 - \lambda_0 C. \quad (48)$$

Let $p_{\mathrm{opt}}(\boldsymbol{x}) \propto \sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}$ by which the equality holds in (48). Then, the Cauchy–Schwarz inequality is tight, and

$$\int \frac{u(\boldsymbol{x})^2}{p_{\mathrm{opt}}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \left(\int \sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}\mathrm{d}\boldsymbol{x}\right)\left(\int \frac{u(\boldsymbol{x})^2}{\sqrt{q(\boldsymbol{x})^2 + \lambda_0 u(\boldsymbol{x})^2}}\mathrm{d}\boldsymbol{x}\right) = h(\lambda_0) = C,$$

so $p_{\mathrm{opt}}$ is feasible and attains the lower bound. Hence, it is optimal.

Finally, when $\lambda_0 = \infty$, we have $(\int u(\boldsymbol{x})\mathrm{d}x)^2 = C$, which means that the constraint is

$$\int \frac{u(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \Big(\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big)^2 \int \frac{\bar{u}(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \Big(\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big)^2\Big(\int \frac{(\bar{u}(\boldsymbol{x}) - p(\boldsymbol{x}))^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} + 1\Big)^2 \leq \Big(\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big)^2$$

for $\bar{u}(\boldsymbol{x}) = u(\boldsymbol{x})/\int u(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$. Thus, the only feasible solution is $p = \bar{u}$. □

# K PROOF OF THEOREM 5

## K.1 ROBUST CONSTRAINT

Our optimization problem is formulated as

$$\min_{p_{\mathrm{n}}:\,\mathrm{p.d.f.}} \int \frac{p_{\mathrm{d}}(\boldsymbol{x})^2}{p_{\mathrm{n}}(\boldsymbol{x})}\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2\mathrm{d}\boldsymbol{x} \quad \text{such that} \quad \forall \boldsymbol{x},\ \frac{p_{\mathrm{d}}(\boldsymbol{x})}{p_{\mathrm{n}}(\boldsymbol{x})}\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma} \leq C'.$$

The proof is completed by substituting $q(\boldsymbol{x}) = p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|/\mathbb{E}_{\mathrm{d}}[\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{X})\|]$ and $u(\boldsymbol{x}) = p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}/C'$ in the following Lemma:

**Lemma 2.** *Let $q(\boldsymbol{x})$ be a probability density and $u(\boldsymbol{x})$ a non-negative function. Consider the optimization problem*

$$\min_{p:\,\mathrm{p.d.f.}} \int \frac{q(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \quad \textit{such that} \quad \forall \boldsymbol{x},\ p(\boldsymbol{x}) \geq u(\boldsymbol{x}).$$

*Suppose that $\int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq 1$. Then the optimal solution is given by*

$$p_{\mathrm{opt}}(\boldsymbol{x}) = \max\{u(\boldsymbol{x}),\, c_0 q(\boldsymbol{x})\},$$

*where $c_0 \geq 0$ is a non-negative constant chosen so that $\int p_{\mathrm{opt}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. Moreover we have $c_0 \in [0, 1]$.*

*Proof of Lemma 2.* Define the function

$$c \mapsto \eta(c) := \int \max\{u(\boldsymbol{x}), cq(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x}.$$

The function $\eta(c)$ is increasing and satisfies $\eta(0) = \int u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq 1$ and $\eta(1) = \int \max\{u(\boldsymbol{x}), q(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x} \geq \int q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. Moreover, by the dominated convergence theorem, $\eta(c)$ is continuous on $[0, 1]$. Hence, by the intermediate value theorem, there exists $c_0 \in [0, 1]$ such that $\eta(c_0) = 1$. If $\eta(0) = 1$, then $p(\boldsymbol{x}) = u(\boldsymbol{x})$ (with $c_0 = 0$) is the unique feasible solution. In the below, we assume $c_0 > 0$.

For a probability density $p(\boldsymbol{x})$, define the functional

$$F(p) := \int \frac{q(\boldsymbol{x})^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \int \Big(\frac{q(\boldsymbol{x})^2}{p(\boldsymbol{x})} + \frac{p(\boldsymbol{x})}{c_0^2}\Big)\mathrm{d}\boldsymbol{x} - \frac{1}{c_0^2}.$$

For any probability density $p(\boldsymbol{x})$ satisfying $p(\boldsymbol{x}) \geq u(\boldsymbol{x})$, we have

$$\frac{q(\boldsymbol{x})^2}{p(\boldsymbol{x})} + \frac{p(\boldsymbol{x})}{c_0^2} \geq \frac{q(\boldsymbol{x})^2}{p_{\mathrm{opt}}(\boldsymbol{x})} + \frac{p_{\mathrm{opt}}(\boldsymbol{x})}{c_0^2}$$

for all $\boldsymbol{x}$. Therefore, $F(p) \geq F(p_{\mathrm{opt}})$ for all feasible $p$, which completes the proof. □

### K.2 ROBUST-INTEGRAL CONSTRAINT

The proof is based on Lemma 1 and is similar as Theorem 4. We first formulate our optimization problem as

$$\min_{p_{\mathrm{n}}:\,\mathrm{p.d.f.}} \int \frac{p_{\mathrm{d}}(\boldsymbol{x})^2}{p_{\mathrm{n}}(\boldsymbol{x})} \|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2 \mathrm{d}\boldsymbol{x} \quad \text{such that} \quad \int \frac{p_{\mathrm{d}}(\boldsymbol{x})}{p_{\mathrm{n}}(\boldsymbol{x})} \|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma} \mathrm{d}\boldsymbol{x} \leq C''.$$

To apply Lemma 1, we next substitute $q(\boldsymbol{x}) = p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|/Z_{\mathrm{n}}$ and $u(\boldsymbol{x}) = \sqrt{p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}}$. By assumption, Case (a) in Lemma 1 is excluded. After the substitution, the left-hand side of the condition in Case (c) of Lemma 1 can be expressed as

$$\int \frac{u(\boldsymbol{x})^2}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = \frac{Z_{\mathrm{n}}}{C''} \int \frac{\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}}{\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|} \mathrm{d}\boldsymbol{x} \geq \frac{Z_{\mathrm{n}}}{C''\kappa_{\max}} \int \|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma-1} \mathrm{d}\boldsymbol{x},$$

where $\kappa_{\max}$ denotes the maximum eigenvalue of $\boldsymbol{I}_{\mathrm{F}}$. Since the right-hand side above diverges in general, Case (c) is also excluded. Hence, the optimal noise distribution is Case (b) and given by

$$\bar{p}_{\mathrm{n}}^{\mathrm{opt}}(\boldsymbol{x};\gamma,\lambda) \propto \begin{cases} \sqrt{p_{\mathrm{d}}(\boldsymbol{x})^2\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2/Z_{\mathrm{n}}^2 + \lambda p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}} & 0 < \lambda < \infty \\ \sqrt{p_{\mathrm{d}}(\boldsymbol{x})}\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma/2} & \lambda = \infty, \end{cases}$$

where $\lambda$ is chosen such that

$$C'' = \left( \int \frac{p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}}{\sqrt{p_{\mathrm{d}}(\boldsymbol{x})^2\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2/Z_{\mathrm{n}}^2 + \lambda p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}}} \mathrm{d}\boldsymbol{x} \right) \left( \int \sqrt{p_{\mathrm{d}}(\boldsymbol{x})^2\|\boldsymbol{I}_{\mathrm{F}}^{-1}\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^2/Z_{\mathrm{n}}^2 + \lambda p_{\mathrm{d}}(\boldsymbol{x})\|\boldsymbol{g}_{\mathrm{m}}(\boldsymbol{x})\|^{\gamma}} \mathrm{d}\boldsymbol{x} \right). \tag{49}$$

## L EXPERIMENTAL DETAILS AND RESULTS WITH CROSS-VALIDATION

For $p_{\mathrm{d}}(\boldsymbol{x})$, $\bar{\delta}(\boldsymbol{x})$ and $p_{\mathrm{n}}(\boldsymbol{x})$, we employed the following probability distributions:

- **Gaussian data ($d = 10$):** The standard normal density $N(\boldsymbol{0}, \boldsymbol{I}_d)$ was used for $p_{\mathrm{d}}(\boldsymbol{x})$, while $N(\boldsymbol{3}, 0.1^2\boldsymbol{I}_d)$ was for $\bar{\delta}(\boldsymbol{x})$ where $N(\boldsymbol{m}, \boldsymbol{C})$ denotes the Gaussian density with the mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$, and $\boldsymbol{I}_d$ is the $d$ by $d$ identity matrix. Then, we estimate the mean vector $\boldsymbol{0}$ in $p_{\mathrm{d}}$ with the following unnormalized model:

$$\log p_{\mathrm{m}}^0(\boldsymbol{x};\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{d}(x_i - \theta_i)^2 - c.$$

  As done in NCE [Gutmann and Hyvärinen, 2010], we used the Gaussian density with the sample mean and sample covariance for $p_{\mathrm{n}}(\boldsymbol{x})$. Estimation errors were measured by $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{0}\|/\sqrt{d}$ where $\widehat{\boldsymbol{\theta}}$ denotes an estimate of $(\theta_1, \ldots, \theta_d)$.
- **Log-normal data ($d = 10$):** The following log-normal density was used for $p_{\mathrm{d}}(\boldsymbol{x})$ and $\bar{\delta}(\boldsymbol{x})$:

$$\prod_{i=1}^{d} \frac{1}{x_i\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log x_i - \mu)^2}{2\sigma^2}\right), \tag{50}$$

  where $x_i > 0$ $(i = 1, \ldots, d)$, and we used $\mu = 1$ and $\mu = 3$ for $p_{\mathrm{d}}(\boldsymbol{x})$ and $\bar{\delta}(\boldsymbol{x})$ respectively, and $\sigma = 1$ in both distributions. An unnormalized model,

$$\log p_{\mathrm{m}}^0(\boldsymbol{x};\boldsymbol{\alpha}) = -\sum_{i=1}^{d}\left\{\log x_i + \frac{1}{2}(\log x_i - \theta_i)^2\right\} - c,$$

  was adopted to estimate the parameter $\mu = 1$ in $p_{\mathrm{d}}(\boldsymbol{x})$, and estimation errors were measured by $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{1}\|/\sqrt{d}$ where $\boldsymbol{1} = (1, \ldots, 1)^\top$. For $p_{\mathrm{n}}(\boldsymbol{y})$, the log-normal density was employed where the empirical estimates were substituted into $\mu$ and $\sigma$ in (50).

- **Poisson data** ($d = 1$)**:** We sampled one-dimensional data $x$ by using Poisson distribution:

$$\frac{\lambda^x e^{-\lambda}}{x!},$$

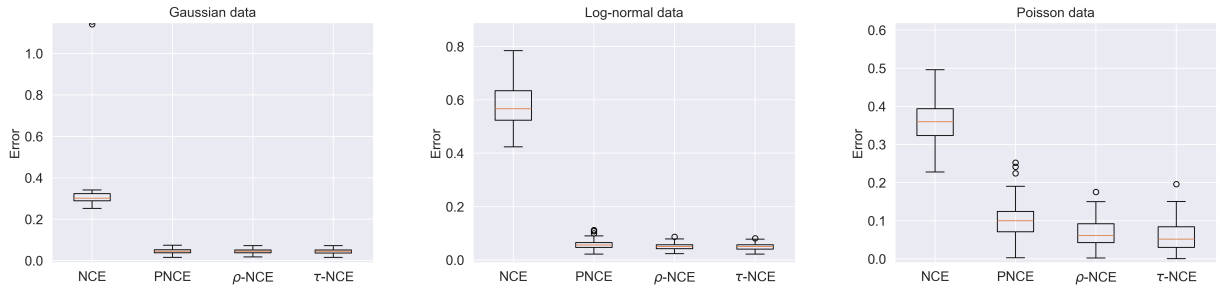
where $\lambda$ is a positive parameter. For data generation, $\lambda = 0.5$ and $\lambda = 5$ were used for $p_{\mathrm{d}}$ and $\bar{\delta}(x)$, respectively. As done in Sasaki and Takenouchi [2024], we employed Poisson distribution with the parameter value twice larger than the empirical estimate of $\lambda$ for $p_{\mathrm{n}}$. To estimate $\lambda$, our unnormalized model is formulated as

$$\log p_{\mathrm{m}}^0(x; \boldsymbol{\alpha}) = x \log \theta + \log x! - c.$$

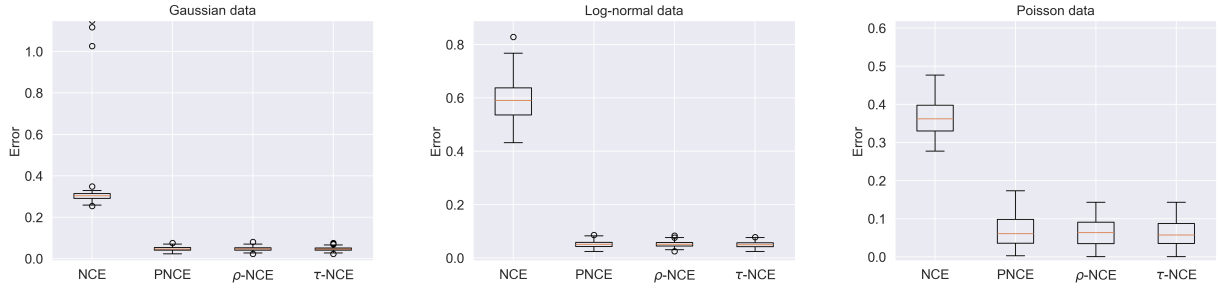
Estimation errors were measured by $|\widehat{\theta} - 0.5|$.

All parameters of unnormalized models were optimized by BFGS.

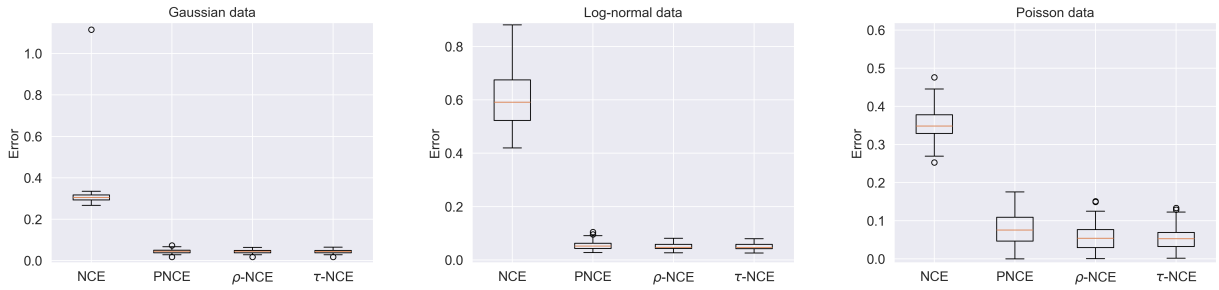
The same experiments as Section 6 were performed with cross-validation in Section 3.1. Figure 7 indicates that our cross-validation method fairly works well on a range of the anchor parameters.



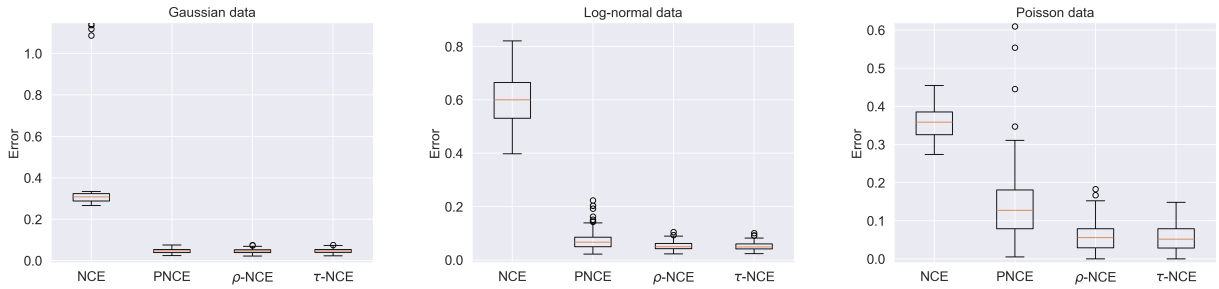

(a) $\beta_0 = \rho_0 = \tau_0 = 0.5$




(b) $\beta_0 = \rho_0 = \tau_0 = 1$




(c) $\beta_0 = \rho_0 = \tau_0 = 3$




(d) $\beta_0 = \rho_0 = \tau_0 = 5$

Figure 7: Estimation errors of NCE, PNCE, $\rho$-NCE and $\tau$-NCE when $(n, n_{\mathrm{y}}, \epsilon) = (1000, 1000, 0.05)$ and anchor parameters ($\beta_0$, $\rho_0$ and $\tau_0$) are changed. The hyper-parameters of PNCE, $\rho$-NCE and $\tau$-NCE are selected based on five-fold cross validation.