# The Optimal Noise in Noise-Contrastive Learning Is Not What You Think


## Abstract

Learning a parametric model of a data distribution is a well-known statistical problem that has seen renewed interest as it is brought to scale in deep learning. Framing the problem as a self-supervised task, where data samples are discriminated from noise samples, is at the core of state-of-the-art methods, beginning with Noise-Contrastive Estimation (NCE). Yet, such contrastive learning requires a good noise distribution, which is hard to specify; domain-specific heuristics are therefore widely used. While a comprehensive theory is missing, it is widely assumed that the optimal noise should in practice be made equal to the data, both in distribution and proportion. This setting underlies Generative Adversarial Networks (GANs) in particular. Here, we empirically and theoretically challenge this assumption on the optimal noise. We show that deviating from this assumption can actually lead to better statistical estimators, in terms of asymptotic variance. In particular, the optimal noise distribution is different from the data's and even from a different family.


## 1 INTRODUCTION

Learning a parametric model of a data distribution is at the core of statistics and machine learning. Once a model is learnt, it can be used to generate new data, to evaluate the likelihood of existing data, or be introspected for meaningful structure such as conditional dependencies between its features. Among an arsenal of statistical methods developed for this problem, Maximum-Likelihood Estimation (MLE) has stood out as the go-to method: given data samples, it evaluates a model's likelihood to have generated them and retains the best fit. However, MLE is limited by the fact that the parametric model has to be properly normalized, which may not be computationally feasible.

In recent years, an alternative has emerged in the form of Noise-Contrastive Estimation (NCE) [Gutmann and Hyvärinen, 2012]: given data samples, it generates noise samples and trains a discriminator to learn the data distribution by constrast. Its supervised formulation, as a binary prediction task, is simple to understand and easy to implement. In fact, NCE can be seen as one of the first, and most fundamental methods of *self-supervised* learning, which has seen a huge increase of interest recently [van den Oord et al., 2018, Chen et al., 2020].

Crucially, NCE can handle unnormalized, i.e. energy-based, models. It has shown remarkable success in Natural Language Processing [Mnih and Teh, 2012, Mikolov et al., 2013] and has spearheaded an entire family of Contrastive Methods [Pihlaja et al., 2010, Gutmann and Hirayama, 2011]. While MLE is known to be optimal in the asymptotic limit of infinite samples, NCE is a popular choice in practice due to its computational advantage. In fact, NCE outperforms Monte Carlo Maximum Likelihood (MLE-MC) [Riou-Durand and Chopin, 2018] - an MLE estimation procedure where normalization is performed by importance sampling.

Nevertheless, NCE's performance is however dependent on two hyperparameters: the choice of noise distribution and the noise-data ratio (or, proportion of noise samples). A natural question follows: what is the optimal choice of noise distribution, and proportion of noise (or, noise-data ratio) for learning the data distribution? There are many heuristics for choosing the noise distribution and ratio in the NCE setting. Conventional wisdom in the related setting of GANs and variants [Goodfellow et al., 2014, Gao et al., 2020] is to set both the proportion and the distribution of noise to be equal to those of the data. The underlying assumption is a game-theoretic notion of optimality: the task of discriminating data from noise is hardest, and therefore most "rewarding", when noise and data are virtually indistinguishable. The noise would then be optimal when the discriminator is no

*Submitted to the 38th Conference on Uncertainty in Artificial Intelligence* (UAI 2022). **To be used for reviewing only**.

longer able to distinguish noise samples from data samples.

However, such an adversarial form of training where a noise generator aims to fool the discriminator suffers from instability and mode-collapse [Dieng et al., 2019, Lorraine et al., 2021]. Furthermore, while the above assumptions (optimal noise equals data) have been supported by numerous empirical successes, it is not clear whether such a choice of noise (distribution and ratio) achieves optimality from a statistical estimation viewpoint. Since NCE is fundamentally motivated by parameter estimation, the optimization of hyperparameters should logically be based on that same framework.

In this work, we propose a principled approach for choosing the optimal noise distribution and ratio, while challenging, both theoretically and empirically, the current practice. In particular, we make the following claims that challenge conventional wisdom:

1. The optimal noise distribution is not the data distribution; in fact, it is of a very different family than the model family.

2. The optimal noise proportion is generally not 50%; the optimal noise-data ratio is not one.

The paper is organized as follows. First, we present NCE and related works in Section 2, as well as the theoretical framework of asymptotic MSE that we use to optimize the NCE estimator. We start Section 3 by empirically showing that the optimal noise distribution is not the data distribution. Our main theoretical results describing the optimal noise distribution are in Section 3.2. Specifically, we analytically provide the optimal noise for NCE, in two interesting limits, and numerically verify how optimal the limits remain outside these limits. We further show empirically that the optimal noise proportion is not 50% either. Finally we discuss the limitations of this work in Section 4 and conclude in Section 5.

**Notation** We denote with $p_d$ a data distribution, $p_n$ a noise distribution, and $(p_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ a parametric family of distributions assumed to contain the data distribution $p_d = p_{\boldsymbol{\theta}^*}$. All distributions are normalized, meaning that the NCE estimator does not consider the normalizing constant as a parameter to be estimated to simplify the analysis: in this setup, NCE can be fairly compared to MLE and the Cramer-Rao bound is well-defined and applicable. The logistic function is denoted by $\sigma(x)$. We will denote by $\nu$ the ratio between the number of noise samples and data samples: $\nu = T_n/T_d$. The notation $\|\cdot\|$ refers to the Euclidean (vector) norm.

# 2 BACKGROUND

## 2.1 DEFINITION OF NCE

Noise-Contrastive Estimation consists in approximating a data distribution $p_d$ by training a discriminator $D(\boldsymbol{x})$ to distinguish data samples $(\boldsymbol{x}_i)_{i\in[1,T_d]} \sim p_d$ from noise samples $(\boldsymbol{x}_i)_{i\in[1,T_n]} \sim p_n$ [Gutmann and Hyvärinen, 2012]. This defines a binary task where $Y = 1$ is the data label and $Y = 0$ is the noise label. The discriminator is optimal when it equals the (Bayes) posterior

$$D(\boldsymbol{x}) = P(Y = 1|X) = \sigma\left(\frac{p_d(\boldsymbol{x})}{\nu p_n(\boldsymbol{x})}\right) \tag{1}$$

i.e. when it learns the density-ratio $\frac{p_d}{p_n}$ [Gutmann and Hyvärinen, 2012, Mohamed and Lakshminarayanan, 2016]. The basic idea in NCE is that replacing in the ratio the data distribution by $p_{\boldsymbol{\theta}}$ and optimizing a discriminator with respect to $\theta$, yields a useful estimator $\hat{\boldsymbol{\theta}}_{\mathrm{NCE}}$ because at the optimum, the model density has to then equal the data density.

Importantly, there is no need for the model to be normalized; the normalization constant (partition function) can be input as an extra parameter, in stark contrast to MLE.

## 2.2 ASYMPTOTIC ANALYSIS

We consider here a very well-known framework to analyze the statistical performance of an estimator. Fundamentally, we are interested in the Mean-Squared Error (MSE), generally defined as

$$\mathbb{E}_{\boldsymbol{\theta}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2] = \mathrm{Var}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) + \mathrm{Bias}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})^2$$

It can mainly be analyzed in the asymptotic regime, with the number of data points $T_d$ being very large. For (asymptotically) unbiased estimators, the estimator's statistical performance is in fact completely characterized by its asymptotic variance (or rather, covariance matrix) because the bias squared is of a lower order for such estimators. The asymptotic variance is classically defined as

$$\boldsymbol{\Sigma} = \lim_{T_d\to\infty} T_d\, \mathbb{E}_{\boldsymbol{\theta}}[(\hat{\boldsymbol{\theta}} - \mathbb{E}_{\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}} - \mathbb{E}_{\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}])^\top] \tag{2}$$

where the estimator is evaluated for each sample size $T_d$ separately. Thus, we use the asymptotic variance compute an asymptotic approximation of the total Mean-Squared Error which we define as

$$\mathrm{MSE} = \frac{1}{T_d}\mathrm{tr}(\boldsymbol{\Sigma}) \ . \tag{3}$$

In the following, we talk about MSE to avoid any confusion regarding the role of bias: we emphasize that the MSE is given by the asymptotic variance since the bias squared is of a lower order (for consistent estimators, and under

some technical constraints). Furthermore, the MSE is always defined in the asymptotic sense as in Eqs. (2)–(3).

When considering normalized distributions, classical statistical theory tells us that the best attainable MSE (among unbiased estimators) is the Cramer-Rao bound, achieved by Maximum-Likelihood Estimation (MLE). This provides a useful baseline, and implies that $\mathrm{MSE}_{\mathrm{NCE}} \geq \mathrm{MSE}_{\mathrm{MLE}}$ necessarily.

In contrast to a classical statistical framework, however, we consider here the case where the bottleneck of the estimator is the computation, while data samples are abundant. This is the case in many modern machine learning applications. The computation can be taken proportional to the total number of data points used, real data and noise samples together, which we denote by $T = T_d + T_n$. Still, the same asymptotic analysis framework can be used.

An asymptotic analysis of NCE has been carried out by Gutmann and Hyvärinen [2012]. The MSE of NCE depends on three design choices (hyperparameters) of the experiment:

- the noise distribution $p_n$
- the noise-data ratio $\nu = T_n/T_d$, from which the noise proportion can be equivalently calculated
- the total number of samples $T = T_d + T_n$, corresponding here to the computational budget

Building on theorem 3 of Gutmann and Hyvärinen [2012], we can write $\mathrm{MSE}_{\mathrm{NCE}}$ as a function of $T$ (not $T_d$) to enforce a finite computational budget, giving

$$\mathrm{MSE}_{\mathrm{NCE}}(T, \nu, p_n) = \frac{\nu+1}{T}\mathrm{tr}(\boldsymbol{I}^{-1} - \frac{\nu+1}{\nu}(\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1})) \tag{4}$$

where $\boldsymbol{m}$ and $\boldsymbol{I}$ are a generalized score mean and covariance, where the integrand is weighted by the term $(1 - D(\boldsymbol{x}))$ involving the optimal discriminator $D(\boldsymbol{x})$:

$$\begin{aligned} \boldsymbol{m} &= \int \boldsymbol{g}(\boldsymbol{x})(1 - D(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x} \\ \boldsymbol{I} &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top(1 - D(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x} \end{aligned} \tag{5}$$

The (Fisher) score vector is the gradient (or derivative in one dimension) of the log of the data distribution with respect to its parameter $\boldsymbol{g}(\boldsymbol{x}) = \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$. Its actual (without the discriminator weight term) mean is null and its covariance is the Fisher Information matrix, written as $I_F = \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p(\boldsymbol{x})d\boldsymbol{x}$ for the rest of the paper.

The question of statistical efficiency of NCE to bridge the gap with MLE therefore becomes to optimize equation 4 with respect to the three hyperparameters.

### 2.3 PREVIOUS WORK

Despite some early results, choosing the best noise distribution to reduce the variance of the NCE estimator remains largely unexplored. Gutmann and Hyvärinen [2012] and Pihlaja et al. [2010] remark that setting $p_n = p_d$ offers a MSE $(1 + \frac{1}{\nu})$ times higher than the Cramer-Rao bound. Therefore, with an infinite budget $T \to \infty$, taking all samples from noise $\nu \to \infty$ brings the $\mathrm{MSE}_{\mathrm{NCE}}$ down to the Cramer-Rao bound.

Motivated by the same goal of improving the statistical efficiency of NCE, Pihlaja et al. [2010], Gutmann and Hirayama [2011] and Uehara et al. [2018] have looked at reducing the variance of NCE. They relax the original NCE objective by writing it as an M-divergence between the distributions $p_d$ and $p_\theta$ [Uehara et al., 2018] or as a Bregman-divergence between the density ratios $\frac{p_d}{\nu p_n}$ and $\frac{p_\theta}{\nu p_n}$. Choosing a divergence boils down to the use of specific non-linearities, which when chosen for the Jensen-Shannon f-divergence leads to the NCE estimator. Pihlaja et al. [2010] numerically explores which non-linearities lead to the lowest MSE, but they explore estimators different from NCE.

More recently, Uehara et al. [2018] show that the asymptotic variance of NCE can be further reduced by using the MLE estimate of the noise parameters obtained from the noise samples, as opposed to the true noise distribution. A similar idea underlies Flow-Contrastive Estimation [Gao et al., 2020]. While this is useful in practice, it does not address the question of finding the optimal noise distribution.

When the noise distribution is fixed, it remains to optimize the noise-data ratio $\nu$ and samples budget $T$. The effect of the samples budget on the NCE estimator is clear: it scales as $\mathrm{MSE}_{\mathrm{NCE}} \propto \frac{1}{T}$. Consequently and remarkably, the optimal noise distribution and noise-data ratio actually do not depend on the budget T. As for the noise-data ratio $\nu$, while Gutmann and Hyvärinen [2012] and Pihlaja et al. [2010] report that NCE reaches Cramer-Rao when both $\nu$ and $T$ tend to infinity, it is of limited practical use due to finite computational resources $T$. In the limit of finite samples, Pihlaja et al. [2010] offers numerical results touching on this matter, although it considers the noise prior is 50% which greatly simplifies the problem as the MSE here becomes linearly dependent on $\nu$.

## 3 OPTIMIZING NOISE IN NCE

In this work we aim to directly optimize the MSE of the original NCE estimator with respect to the noise distribution and noise-data ratio. Analytical optimization of the $\mathrm{MSE}_{\mathrm{NCE}}$ with respect to the noise distribution $p_n$ or ratio $\nu$ is a difficult task: both terms appear nonlinearly within the integrands. Even in the simple case where the data follows a one-dimensional Gaussian distribution parameterized by

variance, as specified in Section 2 of the Supplementary Material, the resulting expression is intractable. This motivates the need for numerical methods.

In the following, we pursue two different strategies for finding the optimal $p_n$. Either $p_n$ can be chosen within the same parametric family as the data distribution (we use the same parametric model for simplicity); this leads to a simple one-dimensional optimization problem (e.g. optimizing a Gaussian mean or variance $\theta$). Or one can relax this assumption and use more flexible "non-parametric" methods, such as a histogram-based expression for $p_n$. In the latter case, assuming the bins of histograms are fixed, one has in practice a higher-dimensional optimization problem with one weight per histogram bin to estimate.

### 3.1 OPTIMIZATION WITHIN THE SAME PARAMETRIC FAMILY

We use here simple data distributions to illustrate the difficulty of finding the optimal distribution. We work with families of a single scalar parameter to make sure that the numerical calculations can be performed exactly.

The data distributions considered from now on are picked among three generative models with a scalar parameter:

(a) a univariate Gaussian parameterized by its mean and whose variance is fixed to 1,

(b) a univariate zero-mean Gaussian parameterized by its variance,

(c) a two-dimensional zero-mean Gaussian parameterized by correlation, i.e. the off-diagonal entries of the covariance matrix. The variables are taken standardized.

While the Gaussian distribution is simple, it is ubiquitous in generative models literature and remains a popular choice in state-of-the-art deep learning algorithms, such as Variational Auto-Encoders (VAEs). Yet, to our knowledge, it remains completely unknown to date how to design the optimal noise to infer the parameters of a Gaussian using NCE.

Assuming the same parametric distribution for the noise as for the data, Figure 1 presents the optimal noise parameter as a function of the data parameter. Details on numerical methods are explained in details below. For the three models above and setting $\nu = 1$, one can observe that the noise parameter systematically differs from the data parameter. They are equal only in the very special case of estimating correlation (case c) for uncorrelated variables. This means that the optimal noise distribution is not equal to the data distribution, even when the noise and the data are restricted to be in the same parametric family of distributions.

Looking more closely, one can notice that the relationship between the optimal noise parameter and the data parameter highly depends on the estimation problem. For model

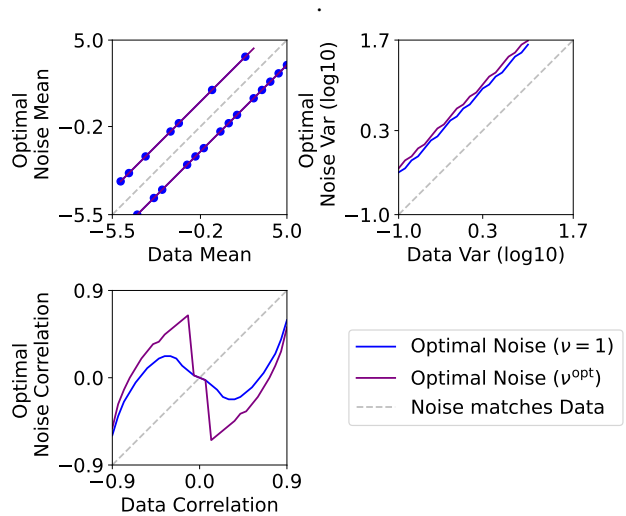



Figure 1: Relationship between the (optimal) noise parameter and the data parameter. (top left) Optimal variance in model (a) as function of the data mean. Note that the noise parameter has two symmetric local minima, given by the individual points, which are joined by a manually drawn line. (top right) Optimal variance in model (b) as function of the data variance. (bottom left) Optimal noise correlation in model (c) as a function of the data correlation.

(a), the optimal noise mean is (randomly) above or below the data mean, while at constant distance (cf. the two local minima of the MSE landscape shown in Section 1 of the Supplementary Material). For model (b), the optimal noise variance is obtained from the data variance by a scaling of $3.84$. This linear relationship is coherent with the symmetry of the problem with respect to the variance parameter. Interestingly for model (c), the optimal noise parameter exhibits a nonlinear relationship to the data parameter: for a very low positive correlation between variables the noise should be negatively correlated, whereas when data variables are strongly correlated, the noise should also be positively correlated.

Having established how different the optimal parametric noise can be, a question naturally follows: what does the optimal, unconstrained noise distribution look like?

### 3.2 THEORY

While the analytical optimization of the noise model is intractable, it is possible to study some limit cases, and by means of Taylor expansions, obtain analytical results which hopefully shed some light to the general behaviour of the estimator even far away from those limits.

In what follows, we study an analytical expression for the optimal noise distribution in three limit cases: (i) when the noise distribution is a (infinitesimal) perturbation of

the data distribution $\frac{p_d}{p_n} \approx 1$; as well as when the noise proportion (ratio) is chosen so that training uses either (ii) all noise samples $\nu \to \infty$ or (iii) all data samples $\nu \to 0$. The following Theorem is proven in Section 4 of the Supplementary Material.

**Theorem 1** *In either of the following two limits:*

*(i) the noise distribution is a (infinitesimal) perturbation of the data distribution $\frac{p_d}{p_n} = 1 + \epsilon(x)$;*

*(ii) in the limit of all noise samples $\nu \to \infty$;*

*the noise distribution minimizing asymptotic MSE is*

$$p_n^{\text{opt}}(\boldsymbol{x}) \propto p_d(\boldsymbol{x}) \|\boldsymbol{I}_F^{-1} \boldsymbol{g}(\boldsymbol{x})\| \ . \tag{6}$$

Interestingly, this is the same as the optimal noise derived by Pihlaja et al. [2010] for another, related estimator (Monte Carlo MLE with Importance Sampling). For example, in the case of estimating Gaussian variance: $p_n^{\text{opt}}(x) \propto \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}} |x^2 - \theta|$ which is highly *non-Gaussian unlike the data distribution*. Similar derivations can be easily done for the cases of Gaussian mean or correlation.

In Section 4 of the Supplementary Material, we further derive a general formula for the gap between the MSE for the typical case $p_n = p_d$ and the optimal $p_n = p_n^{\text{opt}}$. It is given by

$$\Delta\text{MSE} = \frac{1}{T} \text{Var}_{x \sim p_d}(\|\boldsymbol{I}_F^{-1} \boldsymbol{g}(\mathbf{x})\|) \ . \tag{7}$$

This quantity seems to be positive for any reasonable distribution, which implies (in the all-noise limit) that the optimal noise cannot be the data distribution $p_d$. Furthermore, we can compute the gap to efficiency in the all noise limit, i.e. between $p_n = p_n^{\text{opt}}$ and the Cramer-Rao lower bound $\Delta_{\text{opt}}\text{MSE} = \frac{1}{T} \mathbb{E}_{x \sim p_d}(\|\boldsymbol{I}_F^{-1} \boldsymbol{g}(\mathbf{x})\|)^2$.

In the third case, the limit of all data, we have the following conjecture:

**Conjecture 1** *In case (iii), the limit of all data samples $\nu \to 0$, the optimal noise distribution is such that it is all concentrated at the set of those $\boldsymbol{\xi}$ which are given by*

$$\arg\max_{\boldsymbol{\xi}} p_d(\boldsymbol{\xi}) \text{tr}\left( (\boldsymbol{g}(\xi) \boldsymbol{g}(\xi)^\top)^{-1} \right)^{-1} \tag{8}$$

This is typically a degenerate distribution since it is concentrated on a low-dimensional manifold, in the sense of a Dirac delta. For a scalar parameter, the function whose maxima are sought is simply $p_d(\boldsymbol{\xi}) \|\boldsymbol{g}(\xi)\|^2$. An informal proof of this conjecture is given in Section 4 of the Supplementary Material. The "proof" is not quite rigorous due to the singularity of the optimal "density", which is why we label this as conjecture only. While this is in apparent contradiction to having the noise distribution's support include the data distribution's, this result can be understood as a first-order approximation of what one should do with few noise data points available.

Specifically, in the case of estimating a Gaussian mean (for unit variance), one obtains that $p_n^{\text{opt}}(x)$ should be concentrated on two points, given in Dirac delta notation as $\delta_{-\sqrt{2}}$ and $\delta_{\sqrt{2}}$. However, the theory above does not say anything about how the probability mass should be distributed to these two points: it can be 50-50 or all on just one point; all such configurations are equally optimal. This is coherent with the two local minima observed for the Gaussian mean in Figure 1 (top-left). Likewise, when estimating its variance, one obtains, as one possible solution, $p_n^{opt}(x) = \frac{1}{2}(\delta_{-\sqrt{5}} + \delta_{\sqrt{5}})$. These distributions are also highly *non-Gaussian unlike the data distribution*.

So far, we have obtained the optimal noise which minimizes the (asymptotic) estimation error $\mathbb{E}\left[\|\hat{\theta}_T - \theta^*\|^2\right] = \frac{1}{T_d}\text{tr}(\Sigma)$ of NCE for the data *parameter*. However, sometimes estimating the parameter is only a means for estimating the data *distribution* — not an end in itself. We therefore consider the (asymptotic) estimation error induced by the NCE estimator $\hat{\theta}_T$ in the distribution space using the Kullback-Leibler divergence $\mathbb{E}\left[\mathcal{D}_{\text{KL}}(p_d, p_{\hat{\theta}_T})\right]$ which is well-known to equal $\frac{1}{2T_d}\text{tr}(\Sigma I_F)$ (shown in Section 5 of the Supplementary Material). We are thus able to obtain the optimal noise for estimating the data *distribution* in cases (i), (ii) and (iii).

**Theorem 2** *In the two limit cases of Theorem 1, the noise distribution minimizing the expected Kullback-Leibler divergence is given by*

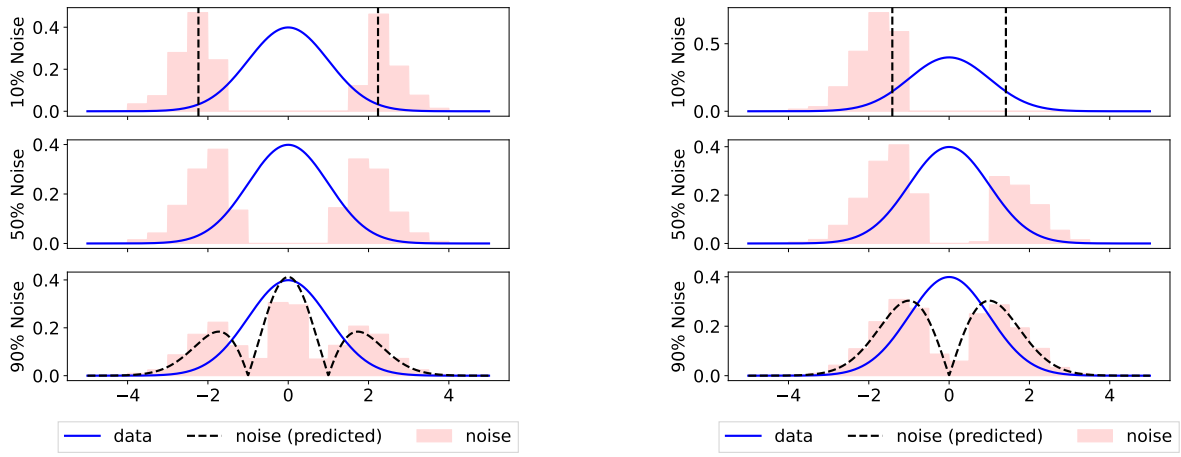
$$p_n^{\text{opt}}(\boldsymbol{x}) \propto p_d(\boldsymbol{x}) \|\boldsymbol{I}_F^{-\frac{1}{2}} \boldsymbol{g}(\boldsymbol{x})\| \ . \tag{9}$$

In the third case, the limit of all data, we have the following conjecture:

**Conjecture 2** *In the limit of Conjecture 1 the noise distribution minimizing the expected Kullback-Leibler divergence is such that it is all concentrated at the set of those $\boldsymbol{\xi}$ which are given by*

$$\arg\max_{\boldsymbol{\xi}} p_d(\boldsymbol{\xi}) \text{tr}\left( (\boldsymbol{g}(\xi) \boldsymbol{g}(\xi)^\top)^{-\frac{1}{2}} \right)^{-1} \tag{10}$$

These optimal noise distributions resemble those from Theorem 1 and Conjecture 1: only the exponent on the Fisher Information matrix changes. This is predictable, as the new cost function $\frac{1}{2T_d}\text{tr}(\Sigma I_F)$ is obtained by scaling with the Fisher Information matrix. More specifically, when the data parameter is scalar, the optimal noises from Theorems 1 and 2 coincide, as the Fisher Information becomes a multiplicative constant; those from Conjectures 1 and 2 do not coincide but are rather similar. The scope of this paper it to investigate the already rich case of a one-dimensional parameter, hence the following focuses on the optimal noise distributions from Theorem 1 and Conjecture 1.




(a) Optimal noise for model (b) (Gaussian variance).

(b) Optimal noise for model (a) (Gaussian mean).

Figure 2: Histogram-based optimal noise distributions. Each row gives a different $\nu$ or noise proportion. The pink bars give the numerical approximations. The theoretical approximation of optimal noise is given by the dashed lines: the all-noise limit in the bottom panel, and the all-data limit in the top panel. In the top panel, the optimal noise is given by single points (Dirac masses) which are chosen symmetric for the purposes of illustration, but as explained in the text, any distribution of probability on those two points is equally optimal.

### 3.3 EXPERIMENTS

We now turn to experiments to validate the theory above. We first describe numerical strategies.

**Numerical Methods** The integrals from Eq. (5) involved in evaluating the asymptotic MSE can be approximated using numerical integration (quadrature) or Monte-Carlo simulations. While both approaches lead to comparable results, quadrature is significantly faster and more precise, especially in low dimension. However, using Monte-Carlo leads to an estimate that is fully differentiable with respect to the parameters of $p_n$.

To tackle the one-dimensional parametric problem, we simply employed quadrature for evaluating the function to optimize over a dense grid and then select the minimum. This appeared as the most computationally efficient strategy and allows for visualizing the MSE landscape reported in Section 1 of the Supplementary Material. In the multi-dimensional non-parametric case, the histogram's weights can be optimized by first-order methods using automatic differentiation.

In the following experiments, the optimization strategy consists in obtaining the gradients of the Monte-Carlo estimate using PyTorch [Paszke et al., 2019] and plugging them into a non-linear conjugate gradient scheme implemented in Scipy [Virtanen et al., 2020]. We chose the conjugate-gradient algorithm as it is deterministic (no residual asymptotic error as with SGD), and as it offered fast convergence. None of the experiments below required more than 100 iterations of conjugate-gradient. Note that for numerical precision, we had to set PyTorch's default to 64-bit floating-point precision. Code is provided in Supplementary Material.

**Results** Figure 2a shows the optimal histogram-based noise distribution for estimating the variance of a zero-mean Gaussian, together with our theoretical predictions (Theorem 1 and Conjecture 1). We can see that our theoretical predictions in the all-data and all-noise limits match such numerical results. It is apparent in Figure 2a that the optimal noise places its mass where the data distribution is high, and where it varies most when $\theta^*$ changes. Furthermore, the noise distribution in the all-data limit has higher mass concentration, which also matches our predictions. Interestingly, in a case not covered by our hypotheses, when there are as many noise samples as data samples i.e. noise proportion of 50% or $\nu = 1$, the optimal noise in Figure 2a (middle) is qualitatively not very different from the limit cases of all data or all noise samples.

Figure 2b gives the same results for the estimation of a Gaussian's mean. The conclusions are similar; in this case, the optimal distributions in the two limits resemble each other even more. It is here important to take into account the indeterminacy of distributing probability mass on the two Diracs, which is coherent with initial experiments in Figure 1 as well as the MSE landscape included in Section 1 of the Supplementary Material.

Figure 3 shows the numerically estimated optimal noise distribution for model (c) using a Gaussian correlation parameter. Here, the distributions are perhaps even more surprising than in previous figures. This can be partly understood by the extremely nonlinear dependence of the optimal noise

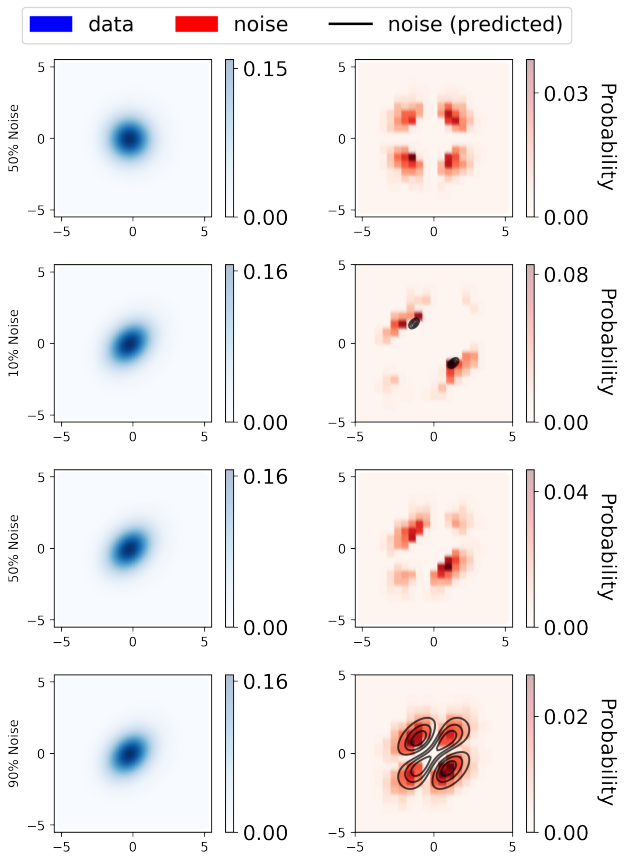



Figure 3: Optimal noise for a 2D Gaussian parameterized by correlation. 2D Gaussian with correlation 0 (top) and 0.3 (bottom three) are considered. Left panel is data density, right panel is the optimal histogram-based noise density. The theoretical approximation of optimal noise is given by the black level lines: the case of Theorem 1 the bottom panel, and the Conjecture 1 in the second panel. Here, the optimal noise in the latter limit is given by a softmax relaxation with temperature 0.01. It makes the choice of placing its mass symmetrically on the single points (Dirac masses), but as explained in the text, any distribution of probability on those two points could be equally optimal.

parameter from the data parameter shown in Fig. 1.

We next ask: how robust to $\nu$ is the analytical noise we derived in these limiting cases? Figure 4 shows the Asymptotic MSE achieved by two noise models, across a range of noise proportions. The first noise model is the optimal noise in the parametric family containing the data distribution $p_n = p_\theta$, optimized for $\nu = 1$, while the second noise model is the optimal analytical noise $p_n^{\text{opt}}$ derived in the all-noise limit, see Eq. (6). They are both compared to the Cramer-Rao lower bound. For all models (a) (b) and (c), the optimal analytical noise $p_n^{\text{opt}}$ (red curve) is empirically useful even far away from the all-noise limit, and across the entire range of noise proportions. In fact, $p_n = p_n^{\text{opt}}$ empirically seems

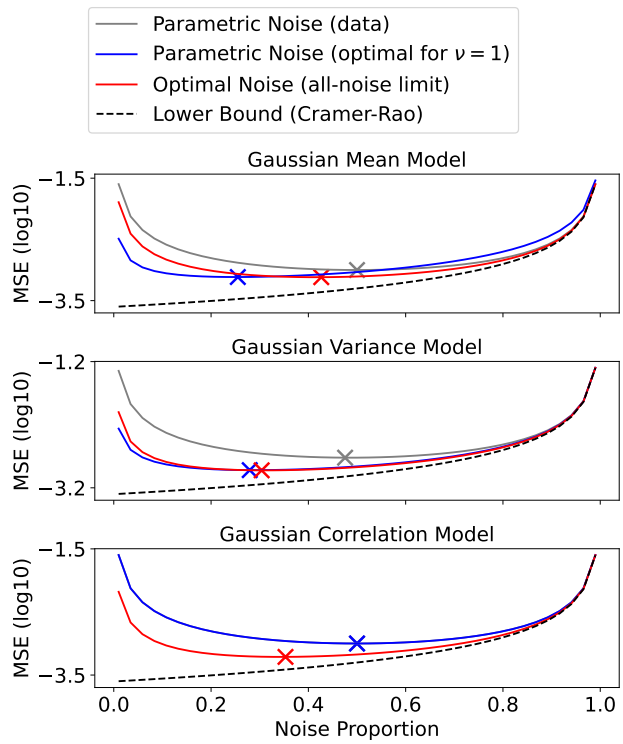



Figure 4: Asymptotic MSE vs. noise proportion. Top panel: Asymptotic MSE vs. noise proportion for model (a) with parameter mean; Middle panel: Asymptotic MSE vs. noise proportion for model (b) with parameter variance; Bottom panel: Asymptotic MSE vs. noise proportion for model (c) with parameter correlation. The parameter in "parametric noise" is the optimal parameter for $\nu = 1$, i.e. for when half the samples are noise and half are data. The "optimal noise" is the approximation given by Theorem 1.

a better choice than using the data distribution $p_n = p_d$, and is (quasi) uniformly equal to or better than a parametric noise $p_n = p_\theta$ optimized for $\nu = 1$.

### 3.4 OPTIMIZING NOISE PROPORTION

Next, we consider optimization of the noise proportion. It is often heuristically assumed that having 50% noise, i.e. $\nu = 1$ is optimal. On the other hand, Pihlaja et al. [2010] provided a general analysis, although it didn't quite answer this question from a practical viewpoint; nor is it compatible with the basic NCE framework of this paper.

In the special case where $p_d = p_n$ we can actually show (see Section 3 of the Supplementary Material) that the optimal noise proportion is $50\%$. This is obtained for a fixed computational budget $T$, as the noise proportion varies between 0 and 1. When this constraint on the budget is relaxed, the optimal noise proportion is $\nu \to \infty$ as in Corollary 7 and Figure 4.d. of Gutmann and Hyvärinen [2012]. The

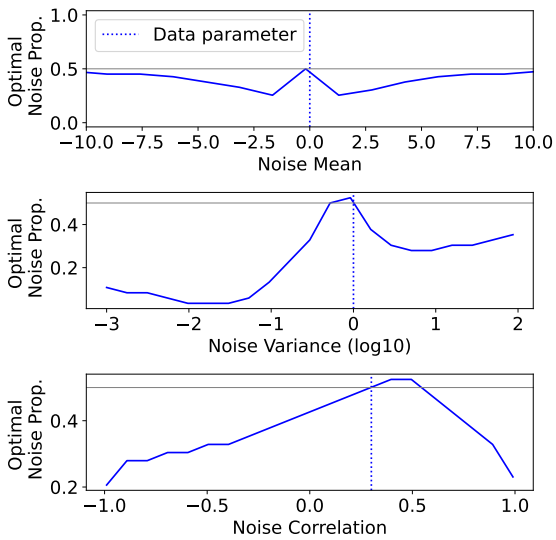



Figure 5: Optimal noise proportion against the noise parameter. Top panel for model (a), Gaussian mean; Middle panel for model (b), Gaussian variance; Bottom panel for model (c), Gaussian correlation.

reciprocal for the theoretical result above does not hold: a noise proportion of $50\%$ does *not* ensure that the noise distribution equals the data's, as shown by counter-examples in Figures 1 and 5.

However, in the general case $p_n \neq p_d$, the optimal proportion is not 50%. We can again look at Figure 4 which analyses the MSE as a function of noise proportion for simple one-parameter families. It is not optimized, in general, at 50%, for the noise distributions considered here. In fact, the parameter of the noise distribution is here optimized for a proportion of 50%, so the results are skewed towards finding that proportion optimal, but still that is not the optimum for most cases.

A closer look at this phenomenon is given by Fig, 5 which shows the optimal noise proportion as a function of a Gaussian's parameter (mean, variance, or correlation). We see that while it is 50% for when the data parameter is used for noise, it is in general less.

## 4 DISCUSSION

We have shown that choosing an optimal noise means choosing a noise distribution that is *different* to the data's. An interesting question is what implications does this have for GANs, which iteratively guide the noise distribution to *match* the data's? Both NCE and GANs in fact solve the binary task of discriminating data from noise. While the optimal discriminator for the binary task recovers the density ratio between noise and data, GANs parameterize the discriminator's denominator (brought closer to the data distribution by adversarial training) while NCE parameterizes the numerator. Hence they do not define the same contrast, though GANs do claim inspiration from NCE [Goodfellow et al., 2014]. Moreover, of course, the goals of the two methods are completely different: GANs do not perform estimation of parameters of a statistical model but focus on the generation of data.

Nevertheless, GAN updates *have* inspired the choice of NCE noise as in Flow-Contrastive Estimation (FCE) by Gao et al. [2020], which parameterizes both the discriminator numerator and discriminator, providing a bridge between NCE and GANs. Results on FCE by Gao et al. [2020] empirically demonstrate that the choice of noise matters: NCE is made quicker by iterative noise updates *à la* GAN, presumably because setting the noise distribution equal to the data's reduces asymptotic variance compared to choosing a generic noise distribution such as the best-matching Gaussian. Noise-updates based on the optimal noise in this paper, could perhaps accelerate convergence even further, avoiding the numerical difficulties of an adversarial game, while still increasing the statistical efficiency.

However, using the optimal noise distributions we present in Section 3.2 can be numerically challenging, especially when the parametric model $p_{\boldsymbol{\theta}}$ is higher-dimensional and unnormalized (e.g. $\theta$ is a dense covariance matrix along with the normalization term as a parameter). Evaluating an optimal noise involves the Fisher score (and therefore access to the very data distribution we seek to estimate) and a Monte-Carlo method may be needed for sampling. We hope that these questions can be resolved in practice by having a relatively simple noise model which is still more statistically efficient than alternatives typically used with NCE, and whose choice is guided by our optimality results.

## 5 CONCLUSION

We studied the choice of optimal design parameters in Noise-Contrastive Estimation. These are essentially the noise distribution and the proportion of noise. We assume that the total number of data points (real data + noise) is fixed due to computational resources, and try to optimize those two hyperparameters. It is easy to show empirically that in stark contrast to what is often assumed, the optimal noise distribution is not the same as the data distribution, thus extending the analysis by Pihlaja et al. [2010]. Our main theoretical results derive the optimal noise distribution in limit cases where either almost all samples to be classified are noise, or almost all samples are real data, or the noise distribution is an (infinitesimal) perturbation of the data distribution. The optimal noise distributions in two of these cases are different but have in common the point of emphasizing parts of the data space where the Fisher score function changes rapidly. We hope these results will improve the performance of NCE in demanding applications.

## References


Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

Adji B. Dieng, Francisco J. R. Ruiz, David M. Blei, and Michalis K. Titsias. Prescribed generative adversarial networks. *ArXiv*, abs/1910.04302, 2019.

Ruiqi Gao, Erik Nijkamp, Diederik P. Kingma, Zhen Xu, Andrew M. Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7515–7525, 2020.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

M. Gutmann and J. Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *UAI*, 2011.

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

Jon Lorraine, David Acuna, Paul Vicol, and D. Duvenaud. Complex momentum for optimization in games. 2021.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *ICML*, 2012.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *ArXiv*, abs/1610.03483, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

M. Pihlaja, M. Gutmann, and A. Hyvärinen. A family of computationally efficient and simple estimators for unnormalized statistical models. In *UAI*, 2010.

Lionel Riou-Durand and Nicolas Chopin. Noise contrastive estimation: Asymptotic properties, formal comparison with MC-MLE. *Electronic Journal of Statistics*, 12:3473–3518, 2018.

Masatoshi Uehara, Takeru Matsuda, and Fumiyasu Komaki. Analysis of noise contrastive estimation from the perspective of asymptotic variance. *ArXiv*, abs/1808.07983, 2018.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, 2018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

# Supplementary Materials

## 1 Visualizations of the MSE landscape

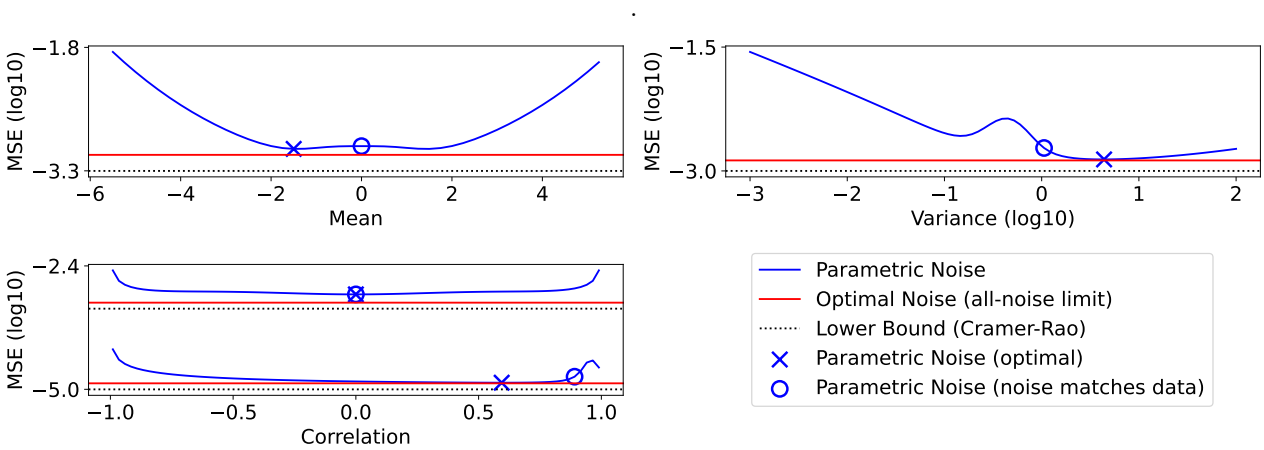



Figure 1: MSE vs. the noise parameter. Top left panel for model (i), Gaussian mean; Top right panel for model (ii), Gaussian variance; Bottom left for model (iii), Gaussian correlation.

We provide visualizations of the MSE landscape of the NCE estimator, when the noise is constrained within a parametric family containing the data.

We draw attention to the two local minima symmetrically placed to the left and to the right of the Gaussian mean. This corroborates the indeterminacies observed in this paper (Conjecture on limit of zero noise), as to where the optimal noise should place its mass for this estimation problem.

## 2 Intractability of the 1D Gaussian case

Suppose the data distribution $p_d$ is a one-dimensional standardized zero-mean Gaussian. The model and noise distributions are of the same family, parameterized by mean and/or variance (we write these together in one model):

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\alpha}}, \quad p_n(x) = \frac{1}{\sqrt{2\pi\beta}} e^{-\frac{1}{2}(\frac{(x-\pi)^2}{\beta}} \qquad x \in \mathbb{R}$$

We can write out the relevant functions, evaluated at $\alpha = 1, \mu = 0$ as the 2D score:

$$\boldsymbol{g}(x) = \begin{pmatrix} \partial_\mu \log p_\theta \\ \partial_\alpha \log p_\theta \end{pmatrix} \Bigg|_{\mu=0,\alpha=1} = \begin{pmatrix} x \\ -1 + x^2 \end{pmatrix}$$

and its "pointwise covariance": $\boldsymbol{g}(x)\boldsymbol{g}(x)^\top = \begin{pmatrix} x^2 & -x + x^3 \\ -x + x^3 & x^4 - x^2 + 1 \end{pmatrix}$

In the following, we consider estimation of variance only. i.e. only the second term in $\boldsymbol{m}$ and the second diagonal term in the Fisher information matrix $I$ . Now we can compute the generalized score mean $m$ and mean of square $I$ as they intervene in the MSE formula for Noise-Contrastive Estimation:

$$m = \int g(x)(1 - D(x))p(x)dx$$

which gives

$$m = -\frac{1}{2\sqrt{2\pi}} \int \left( e^{\frac{-x^2}{2}} \frac{1}{1 + \frac{1}{\nu}\sqrt{\beta} e^{\frac{-x^2}{2}(1-\frac{1}{\beta})}} \right) dx + \frac{1}{2\sqrt{2\pi}} \int x^2 \left( e^{\frac{-x^2}{2}} \frac{1}{1 + \frac{1}{\nu}\sqrt{\beta} e^{\frac{-x^2}{2}(1-\frac{1}{\beta})}} \right) dx$$

and

$$I = \int g(x)^2(1 - D(x))p(x)dx$$

which gives

$$I = \frac{1}{4\sqrt{2\pi}} \int x^4 \left( e^{\frac{-x^4}{2}} \frac{1}{1 + \frac{1}{\nu}\sqrt{\beta} e^{\frac{-x^2}{2}(1-\frac{1}{\beta})}} \right) dx - \frac{1}{2\alpha^3\sqrt{2\pi}} \int x^2 \left( e^{\frac{-x^2}{2}} \frac{1}{1 + \frac{1}{\nu}\sqrt{\beta} e^{\frac{-x^2}{2}(1-\frac{1}{\beta})}} \right) dx$$
$$+ \frac{1}{4\sqrt{2\pi}} \int \left( e^{\frac{-x^2}{2}} \frac{1}{1 + \frac{1}{\nu}\sqrt{\beta} e^{\frac{-x^2}{2}(1-\frac{1}{\beta})}} \right) dx$$

We see that even in a simple 1D Gaussian setting, evaluating the asymptotic MSE of the Noise-Contrastive Estimator is untractable in closed-form, given the integrals in $I$, where the integrand includes the product of a Gaussian density with the logistic function compounded by the Gaussian density, further multiplied by monomials. While here we considered the case of variance, the intractability is seen even in the case of the mean. Optimizing the asymptotic MSE with respect to $\beta$ and $\pi$ (noise distribution) or $\nu$ (identifiable to the noise proportion) yields similarly intractable integrals.

## 3 Optimal Noise Proportion when the Noise Distribution matches the Data Distribution: Proof

We wish to minimize the MSE given by equation 4

$$\mathrm{MSE}_{\mathrm{NCE}}(T, \nu, p_n) = \frac{\nu+1}{T}\mathrm{tr}(\boldsymbol{I}^{-1} - \frac{\nu+1}{\nu}(\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1}))$$

when $p_n = p_d$. In that case,

$$D(\boldsymbol{x}) = \frac{p_d}{p_d + \nu p_n}(\boldsymbol{x}) = \frac{p_d}{p_d + \nu p_d}(\boldsymbol{x}) = \frac{1}{1+\nu}$$

and the integrals involved become

$$\begin{aligned}\boldsymbol{m} &= \int \boldsymbol{g}(\boldsymbol{x})(1 - D(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{\nu}{1+\nu}\int \boldsymbol{g}(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \\ &= 0\end{aligned}$$

given the score has zero mean, and

$$\begin{aligned}\boldsymbol{I} &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top(1 - D(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{\nu}{1+\nu}\int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{\nu}{1+\nu}\boldsymbol{I}_F \quad .\end{aligned}$$

The objective function thus reduces to

$$\mathrm{MSE}_{\mathrm{NCE}}(T, \nu, p_n) = \frac{\nu+1}{T}\mathrm{tr}(\boldsymbol{I}^{-1}) = \frac{(\nu+1)^2}{\nu T}\mathrm{tr}(\boldsymbol{I}_F^{-1}) \propto \frac{(\nu+1)^2}{\nu} \quad .$$

The derivative with respect to $\nu$ is proportional to $\frac{1}{\nu^2} - 1$ and is null when $\nu = 1$ so when the noise proportion is 50%.

# 4 Optimal Noise for Estimating a Parameter: Proofs

We here prove the theorem and conjecture for the optimal noise distribution in three limit cases $\nu \to 0$ (all data samples), $\nu \to \infty$ (all noise samples), and $\frac{p_d}{p_n}(.) = 1 + \epsilon(.)$ as $\epsilon(.) \to 0$ (noise distribution is an infinitesimal perturbation of the data distribution).

The goal is to optimize the $\mathrm{MSE}_{\mathrm{NCE}}(T, \nu, p_n)$ with respect to the noise distribution $p_n$, where

$$\mathrm{MSE}_{\mathrm{NCE}}(T, \nu, p_n) = \frac{\nu+1}{T}\mathrm{tr}(\boldsymbol{I}^{-1} - \frac{\nu+1}{\nu}(\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1})) \tag{1}$$

where the integrals

$$\boldsymbol{m} = \int \boldsymbol{g}(\boldsymbol{x})(1 - D(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x}$$
$$\boldsymbol{I} = \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top(1 - D(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x}$$

depend non-linearly on $p_n$ via the optimal discriminator:

$$1 - D(\boldsymbol{x}) = \frac{\nu p_n(\boldsymbol{x})}{p_d(\boldsymbol{x}) + \nu p_n(\boldsymbol{x})}$$

The general proof structure is:

- Perform a Taylor expansion of $1 - D(\boldsymbol{x})$ in the $\nu \to 0$ or $\nu \to \infty$ limit
- Plug into the integrals $\boldsymbol{m}$, $\boldsymbol{I}$ and evaluate them (up to a certain order)
- Perform a Taylor expansion of $\boldsymbol{I}^{-1}$ (up to a certain order)
- Evaluate the $\mathrm{MSE}_{\mathrm{NCE}}$ (up to a certain order)

- Optimize the $\mathrm{MSE}_{\mathrm{NCE}}$ w.r.t. $p_n$
- Compute the MSE gaps at optimality

**Theorem 1** *In cases:*
*(i) when the noise distribution is a (infinitesimal) perturbation of the data distribution* $\frac{p_d}{p_n} = 1 + \epsilon(x)$*;*
*(ii) in the limit of all noise samples* $\nu \to \infty$*;*
*the optimal noise distributions match. They are given by:*

$$p_n^{\text{opt}}(\boldsymbol{x}) \propto p_d(\boldsymbol{x}) \|\boldsymbol{I}_F^{-1} \boldsymbol{g}(\boldsymbol{x})\| \ . \tag{2}$$

*Proof:* case of $\nu \to \infty$

We start with a change of variables $\gamma = \frac{1}{\nu} \to 0$ to bring us to a zero-limit.

The MSE in terms of our new variable $\gamma = \frac{1}{\nu}$ can be written as:

$$\text{MSE}_{\text{NCE}}(T, \gamma, p_n) = \frac{\gamma + 1}{\gamma T} \text{tr}(\boldsymbol{I}^{-1}) - \frac{(\gamma+1)^2}{T\gamma} \text{tr}(\boldsymbol{I}^{-1} \boldsymbol{m}\boldsymbol{m}^\top \boldsymbol{I}^{-1}) \tag{3}$$

$$= \left(\gamma^{-1} T^{-1} + \gamma^0 T^{-1}\right) \text{tr}(\boldsymbol{I}^{-1}) - \left(\gamma^{-1} T^{-1} + \gamma^0 2T^{-1} + \gamma^1 T^{-1}\right) \text{tr}(\boldsymbol{I}^{-1} \boldsymbol{m}\boldsymbol{m}^\top \boldsymbol{I}^{-1}) \tag{4}$$

Given the term up until $\gamma^-1$ in the MSE, we will use Taylor expansions up to order 2 throughout the proof, in anticipation that the MSE will be expanded until order 1.

- Taylor expansion of the discriminator

$$1 - D(\boldsymbol{x}) = \frac{\nu p_n(\boldsymbol{x})}{p_d(\boldsymbol{x}) + \nu p_n(\boldsymbol{x})} = \frac{1}{1 + \gamma \frac{p_d}{p_n}(\boldsymbol{x})} = 1 - \gamma \frac{p_d}{p_n}(\boldsymbol{x}) + \gamma^2 \frac{p_d^2}{p_n^2}(\boldsymbol{x}) + \circ(\gamma^2)$$

- Evaluating the integrals $\boldsymbol{m}$, $\boldsymbol{I}$

$$\boldsymbol{m} = \int \boldsymbol{g}(\boldsymbol{x}) p_d(\boldsymbol{x}) \Big(1 - D(\boldsymbol{x})\Big) d\boldsymbol{x} = \int \boldsymbol{g}(\boldsymbol{x}) p_d(\boldsymbol{x}) \left(1 - \gamma \frac{p_d}{p_n}(\boldsymbol{x}) + \gamma^2 \frac{p_d^2}{p_n^2}(\boldsymbol{x}) + \circ(\gamma^2)\right) d\boldsymbol{x}$$
$$= \boldsymbol{m}_F - \gamma \boldsymbol{a} + \gamma^2 \boldsymbol{b} + \circ(\gamma^2)$$

where the Fisher-score mean $\boldsymbol{m}_F$ is null and we use shorthand notations $a$ and $b$ for the remaining integrals:

$$\boldsymbol{m}_F = \int \boldsymbol{g}(\boldsymbol{x}) p_d(\boldsymbol{x}) d\boldsymbol{x} = 0$$
$$\boldsymbol{a} = \int \boldsymbol{g}(\boldsymbol{x}) \frac{p_d^2}{p_n}(\boldsymbol{x}) d\boldsymbol{x}$$
$$\boldsymbol{b} = \int \boldsymbol{g}(\boldsymbol{x}) \frac{p_d^3}{p_n^2}(\boldsymbol{x}) d\boldsymbol{x} \ .$$

Similarly,

$$\boldsymbol{I} = \int \boldsymbol{g}(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x})^\top p_d(\boldsymbol{x}) \Big(1 - D(\boldsymbol{x})\Big) d\boldsymbol{x} = \int \boldsymbol{g}(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x})^\top p_d(\boldsymbol{x}) \left(1 - \gamma \frac{p_d}{p_n}(\boldsymbol{x}) + \gamma^2 \frac{p_d^2}{p_n^2}(\boldsymbol{x}) + \circ(\gamma^2)\right) d\boldsymbol{x}$$
$$= \boldsymbol{I}_F - \gamma \boldsymbol{A} + \gamma^2 \boldsymbol{B} + \circ(\gamma^2)$$

where the Fisher-score covariance (Fisher information) is $\boldsymbol{I}_F$ and we use shorthand notations $A$ and $B$ for the remaining integrals:

$$\boldsymbol{I}_F = \int \boldsymbol{g}(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x})^\top p_d(\boldsymbol{x}) d\boldsymbol{x}$$
$$\boldsymbol{A} = \int \boldsymbol{g}(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x})^\top \frac{p_d^2}{p_n}(\boldsymbol{x}) d\boldsymbol{x}$$
$$\boldsymbol{B} = \int \boldsymbol{g}(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x})^\top \frac{p_d^3}{p_n^2}(\boldsymbol{x}) d\boldsymbol{x} \ .$$

- Taylor expansion of $\boldsymbol{I}^{-1}$

$$\begin{aligned}
\boldsymbol{I}^{-1} &= \Big(\boldsymbol{I}_F - \gamma\boldsymbol{A} + \gamma^2\boldsymbol{B} + \circ(\gamma^2)\Big)^{-1} \\
&= \Big(\boldsymbol{I}_F(\mathbf{Id} - \gamma\boldsymbol{I}_F^{-1}\boldsymbol{A} + \gamma^2\boldsymbol{I}_F^{-1}\boldsymbol{B}) + \circ(\gamma^2)\Big)^{-1} \\
&= \boldsymbol{I}_F^{-1}\Big(\mathbf{Id} - \gamma\boldsymbol{I}_F^{-1}\boldsymbol{A} + \gamma^2\boldsymbol{I}_F^{-1}\boldsymbol{B}\Big)^{-1} + \circ(\gamma^2) \\
&= \boldsymbol{I}_F^{-1}\Big(\mathbf{Id} + \gamma\boldsymbol{I}_F^{-1}\boldsymbol{A} + \gamma^2((\boldsymbol{I}_F^{-1}\boldsymbol{A})^2 - \boldsymbol{I}_F^{-1}\boldsymbol{B}) + \circ(\gamma^2)\Big) + \circ(\gamma^2) \\
&= \boldsymbol{I}_F^{-1} + \gamma\boldsymbol{I}_F^{-2}\boldsymbol{A} + \gamma^2(\boldsymbol{I}_F^{-1}(\boldsymbol{I}_F^{-1}\boldsymbol{A})^2 - \boldsymbol{I}_F^{-2}\boldsymbol{B}) + \circ(\gamma^2)
\end{aligned}$$

- Evaluating the $\mathrm{MSE}_{\mathrm{NCE}}$

$$\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1} = \gamma^2\boldsymbol{I}_F^{-1}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{I}_F^{-1} + \circ(\gamma^2)$$

by plugging in the Taylor expansions of $\boldsymbol{I}^{-1}$ and $\boldsymbol{m}$ and retaining only terms up to the second order. Hence, the second term of the MSE without the trace is

$$\begin{aligned}
&\Big(\gamma^{-1}T^{-1} + \gamma^0 2T^{-1} + \gamma^1 T^{-1}\Big)\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1} \\
&= \gamma\frac{1}{T}\boldsymbol{I}_F^{-1}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{I}_F^{-1} + \circ(\gamma)
\end{aligned}$$

and the first term of the MSE without the trace is

$$\begin{aligned}
&\Big(\gamma^{-1}T^{-1} + \gamma^0 T^{-1}\Big)(\boldsymbol{I}^{-1}) \\
&= \Big(\gamma^{-1}T^{-1} + \gamma^0 T^{-1}\Big)\Big(\boldsymbol{I}_F^{-1} + \gamma\boldsymbol{I}_F^{-2}\boldsymbol{A} + \gamma^2(\boldsymbol{I}_F^{-1}(\boldsymbol{I}_F^{-1}\boldsymbol{A})^2 - \boldsymbol{I}_F^{-2}\boldsymbol{B}) + \circ(\gamma^2)\Big) \\
&= \gamma^{-1}\frac{1}{T}\boldsymbol{I}_F^{-1} + \gamma^0\frac{1}{T}(\boldsymbol{I}_F^{-2}\boldsymbol{A} + \boldsymbol{I}_F^{-1}) + \gamma^1\frac{1}{T}[\boldsymbol{I}_F^{-1}(\boldsymbol{I}_F^{-1}\boldsymbol{A})^2 - \boldsymbol{I}_F^{-2}\boldsymbol{B} + \boldsymbol{I}_F^{-2}\boldsymbol{A}] + \circ(\gamma) \ .
\end{aligned}$$

Subtracting the second term from the first term and applying the trace, we finally write the MSE:

$$\mathrm{MSE}_{\mathrm{NCE}} = \mathrm{tr}(\gamma^{-1}\frac{1}{T}\boldsymbol{I}_F^{-1} + \gamma^0\frac{1}{T}(\boldsymbol{I}_F^{-2}\boldsymbol{A} + \boldsymbol{I}_F^{-1}) + \gamma^1\frac{1}{T}[\boldsymbol{I}_F^{-1}(\boldsymbol{I}_F^{-1}\boldsymbol{A})^2 - \boldsymbol{I}_F^{-2}\boldsymbol{B} + \boldsymbol{I}_F^{-2}\boldsymbol{A} - \boldsymbol{I}_F^{-1}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{I}_F^{-1}] + \circ(\gamma)) \tag{5}$$

- Optimize the $\mathrm{MSE}_{\mathrm{NCE}}$ w.r.t. $p_n$

To optimize w.r.t. $p_n$, we need only keep the two first orders of the $\mathrm{MSE}_{\mathrm{NCE}}$, which depends on $p_n$ only via the term $\mathrm{tr}(\boldsymbol{I}_F^{-2}\boldsymbol{A}) = \int \|\boldsymbol{I}_F^{-1}\boldsymbol{g}(x)\|^2\frac{p_d^2}{p_n}(x)d\boldsymbol{x}$. Hence, we need to optimize

$$J(p_n) = \frac{1}{T}\int \|I_F^{-1}\boldsymbol{g}(\boldsymbol{x})\|^2\frac{p_d^2}{p_n}(\boldsymbol{x})d\boldsymbol{x} \tag{6}$$

with respect to $p_n$. We compute the variational (Fréchet) derivative together with the Lagrangian of the constraint $\int p_n(\boldsymbol{x}) = 1$ (with $\lambda$ denoting the Lagrangian multiplier) to obtain

$$\delta_{p_n}J = -\|\boldsymbol{I}_F^{-1}\boldsymbol{g}\|^2\frac{p_d^2}{p_n^2} + \lambda \ . \tag{7}$$

Setting this to zero and taking into account the non-negativity of $p_n$ gives

$$p_n(\boldsymbol{x}) = \|\boldsymbol{I}_F^{-1}\boldsymbol{g}(x)\|p_d(\boldsymbol{x})/Z \tag{8}$$

where $Z = \int \|\boldsymbol{I}_F^{-1}\boldsymbol{g}(x)\|p_d(\boldsymbol{x})d\boldsymbol{x}$ is the normalization constant. This is thus the optimal noise distribution, as a first-order approximation.

- Compute the MSE gaps at optimality

  Plugging this optimal $p_n$ into the formula of $\mathrm{MSE}_{\mathrm{NCE}}$ and subtracting the Cramer-Rao lower bound, we get:

$$\begin{aligned}\Delta_{\mathrm{opt}}\mathrm{MSE}_{\mathrm{NCE}} &= \mathrm{MSE}_{\mathrm{NCE}}(p_n = p_n^{\mathrm{opt}}) - \mathrm{MSE}_{\mathrm{Cramer-Rao}} \\ &= \frac{1}{T}\left(\int \|I_F^{-1}\boldsymbol{\psi}\| p_d\right)^2 .\end{aligned}$$

  This is interesting to compare with the case where the noise distribution is the data distribution, which gives

$$\begin{aligned}\Delta_{\mathrm{data}}\mathrm{MSE}_{\mathrm{NCE}} &= \mathrm{MSE}_{\mathrm{NCE}}(p_n = p_d) - \mathrm{MSE}_{\mathrm{Cramer-Rao}} \\ &= \frac{1}{T}\int \|I_F^{-1}\boldsymbol{\psi}\|^2 p_d\end{aligned}$$

  where the squaring is in a different place. In fact, we can compare these two quantities by the Cauchy-Schwartz inequality, or simply the fact that

$$\begin{aligned}\Delta\mathrm{MSE}_{\mathrm{NCE}} &= \Delta_{\mathrm{data}}\mathrm{MSE}_{\mathrm{NCE}} - \Delta_{\mathrm{opt}}\mathrm{MSE}_{\mathrm{NCE}} \\ &= \mathrm{MSE}_{\mathrm{NCE}}(p_n = p_d) - \mathrm{MSE}_{\mathrm{NCE}}(p_n = p_n^{\mathrm{opt}}) \\ &= \frac{1}{T}\mathrm{Var}_{X\sim p_d}\{\|I_F^{-1}\boldsymbol{g}(\mathbf{X})\|\}\end{aligned}$$

  This implies that the two MSEs, when when the noise distribution is either $p_n^{\mathrm{opt}}$ or $p_d$, can be equal only if $\|I_F^{-1}\boldsymbol{g}(.)\|$ is constant in the support of $p_d$. This does not seem to be possible for any reasonable distribution.

*Proof:* case where $p_n \approx p_d$

We consider the limit case where $\frac{p_d}{p_n}(\boldsymbol{x}) = 1 + \epsilon(\boldsymbol{x})$ with $|\epsilon(\boldsymbol{x}) - 0| < \epsilon_{\max} \quad \forall \boldsymbol{x}$.

Note that in order to use Taylor expansions for terms containing $\epsilon(\boldsymbol{x})$ in an integral, we assume for any integrand $h(\boldsymbol{x})$ that $\int h(\boldsymbol{x})\epsilon(\boldsymbol{x})d\boldsymbol{x} \approx \epsilon \int h(\boldsymbol{x})d\boldsymbol{x}$, where $\epsilon$ would be a constant.

- Taylor expansion of the discriminator

$$1 - D(\boldsymbol{x}) = \frac{\nu p_n(\boldsymbol{x})}{p_d(\boldsymbol{x}) + \nu p_n(\boldsymbol{x})} = \frac{1}{1 + \frac{1}{\nu} + \frac{p_d}{p_n}(\boldsymbol{x})} = \frac{1}{1 + \frac{1}{\nu} + \frac{1}{\nu}\epsilon(\boldsymbol{x})} = \frac{\nu}{1+\nu}\epsilon^0(\boldsymbol{x}) - \frac{\nu}{(1+\nu)^2}\epsilon^1(\boldsymbol{x}) + \frac{\nu}{(1+\nu)^3}\epsilon^2(\boldsymbol{x}) + \circ(\epsilon^2)$$

- Evaluating the integrals $\boldsymbol{m}$, $\boldsymbol{I}$

$$\begin{aligned}\boldsymbol{m} &= \int \boldsymbol{g}(\boldsymbol{x})p_d(\boldsymbol{x})\Big(1 - D(\boldsymbol{x})\Big)d\boldsymbol{x} = \int \boldsymbol{g}(\boldsymbol{x})p_d(\boldsymbol{x})\Big(\frac{\nu}{1+\nu}\epsilon^0(\boldsymbol{x}) - \frac{\nu}{(1+\nu)^2}\epsilon^1(\boldsymbol{x}) + \frac{\nu}{(1+\nu)^3}\epsilon^2(\boldsymbol{x}) + \circ(\epsilon^2)\Big)d\boldsymbol{x} \\ &= \frac{\nu}{1+\nu}\boldsymbol{m}_F - \frac{\nu}{(1+\nu)^2}\boldsymbol{a}(\epsilon) + \frac{\nu}{(1+\nu)^3}\boldsymbol{b}(\epsilon^2) + \circ(\epsilon^3)\end{aligned}$$

where the Fisher-score mean $\boldsymbol{m}_F$ is null and we use shorthand notations $a$ and $b$ for the remaining integrals:

$$\begin{aligned}\boldsymbol{m}_F &= \int \boldsymbol{g}(\boldsymbol{x})p_d(\boldsymbol{x})d\boldsymbol{x} = 0 \\ \boldsymbol{a}(\epsilon) &= \int \boldsymbol{g}(\boldsymbol{x})p_d\epsilon(\boldsymbol{x})d\boldsymbol{x} \\ \boldsymbol{b}(\epsilon^2) &= \int \boldsymbol{g}(\boldsymbol{x})p_d\epsilon^2(\boldsymbol{x})d\boldsymbol{x} \ .\end{aligned}$$

Similarly,

$$\begin{aligned}\boldsymbol{I} &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p_d(\boldsymbol{x})\Big(1 - D(\boldsymbol{x})\Big)d\boldsymbol{x} = \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p_d(\boldsymbol{x})\Big(\frac{\nu}{1+\nu}\epsilon^0(\boldsymbol{x}) - \frac{\nu}{(1+\nu)^2}\epsilon^1(\boldsymbol{x}) + \frac{\nu}{(1+\nu)^3}\epsilon^2(\boldsymbol{x}) + \circ(\epsilon^2)\Big)d\boldsymbol{x} \\ &= \frac{\nu}{1+\nu}\boldsymbol{I}_F - \frac{\nu}{(1+\nu)^2}\boldsymbol{A}(\epsilon) + \frac{\nu}{(1+\nu)^3}\boldsymbol{B}(\epsilon^2) + \circ(\epsilon^3)\end{aligned}$$

where the Fisher-score covariance (Fisher information) is $\boldsymbol{I}_F$ and we use shorthand notations $A$ and $B$ for the remaining integrals:

$$\begin{aligned}\boldsymbol{I}_F &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p_d(\boldsymbol{x})d\boldsymbol{x} \\ \boldsymbol{A}(\epsilon) &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p_d\epsilon(\boldsymbol{x})d\boldsymbol{x} \\ \boldsymbol{B}(\epsilon^2) &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p_d\epsilon^2(\boldsymbol{x})d\boldsymbol{x} \ .\end{aligned}$$

- Taylor expansion of $\boldsymbol{I}^{-1}$

$$\begin{aligned}\boldsymbol{I}^{-1} &= \left(\frac{\nu}{1+\nu}\boldsymbol{I}_F - \frac{\nu}{(1+\nu)^2}\boldsymbol{A}(\epsilon) + \frac{\nu}{(1+\nu)^3}\boldsymbol{B}(\epsilon^2) + \circ(\epsilon^3)\right)^{-1} \\ &= \frac{1+\nu}{\nu}\boldsymbol{I}_F^{-1} + \frac{1}{\nu}\boldsymbol{I}_F^{-2}\boldsymbol{A}(\epsilon) + \frac{\nu}{1+\nu}\boldsymbol{I}_F^{-2}\big(\boldsymbol{I}_F^{-1}\boldsymbol{A}^2(\epsilon) - \boldsymbol{B}(\epsilon^2)\big) + \circ(\epsilon^3)\end{aligned}$$

- Evaluating the $\mathrm{MSE}_{\mathrm{NCE}}$

$$\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1} = \frac{1}{(1+\nu)^2}\boldsymbol{I}_F^{-1}\boldsymbol{a}(\epsilon)\boldsymbol{a}(\epsilon)^\top\boldsymbol{I}_F^{-1} + \circ(\epsilon^3)$$

by plugging in the Taylor expansions of $\boldsymbol{I}^{-1}$ and $\boldsymbol{m}$ and retaining only terms up to the second order. Finally, the MSE becomes:

$$\begin{aligned}\mathrm{MSE}_{\mathrm{NCE}}(T,\nu,p_n) &= \frac{\nu+1}{T}\mathrm{tr}(\boldsymbol{I}^{-1} - \frac{\nu+1}{\nu}(\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1})) \\ &= \mathrm{tr}\Big(\frac{(1+\nu)^2}{T\nu}\boldsymbol{I}_F^{-1} + \frac{1+\nu}{T\nu}\boldsymbol{I}_F^{-2}\boldsymbol{A}(\epsilon) + \frac{1}{T\nu}\big(\boldsymbol{I}_F^{-3}\boldsymbol{A}^2(\epsilon) - \boldsymbol{I}_F^{-2}\boldsymbol{B}(\epsilon^2) - \boldsymbol{I}_F^{-1}\boldsymbol{a}(\epsilon)\boldsymbol{a}(\epsilon)^\top\boldsymbol{I}_F^{-1}\big)\Big) + \circ(\epsilon^3)\end{aligned}$$

- Optimize the $\mathrm{MSE}_{\mathrm{NCE}}$ w.r.t. $p_n$

  To optimize w.r.t. $p_n$, we need only keep the $\mathrm{MSE}_{\mathrm{NCE}}$ up to order 1, which depends on $p_n$ only via the term

$$\mathrm{tr}(\boldsymbol{I}_F^{-2}\boldsymbol{A}(\epsilon)) = \mathrm{tr}\Big(\boldsymbol{I}_F^{-2}\big(\int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top\frac{p_d^2}{p_n}(x)d\boldsymbol{x} - \boldsymbol{I}_F\big)\Big)$$

  . where we unpacked $p_n$ from $\epsilon = \frac{p_d}{p_n} - 1$. Hence, we need to optimize

$$J(p_n) = \frac{1}{T}\int \|I_F^{-1}\boldsymbol{g}(\boldsymbol{x})\|^2\frac{p_d^2}{p_n}(\boldsymbol{x})d\boldsymbol{x} \tag{9}$$

  with respect to $p_n$. This was already done in the all-noise limit $\nu \to \infty$ and yielded

$$p_n(\boldsymbol{x}) = \|\boldsymbol{I}_F^{-1}\boldsymbol{g}(x)\|p_d(\boldsymbol{x})/Z \tag{10}$$

  where $Z = \int \|\boldsymbol{I}_F^{-1}\boldsymbol{g}(x)\|p_d(\boldsymbol{x})d\boldsymbol{x}$ is the normalization constant. This is thus the optimal noise distribution, as a first-order approximation.

In the third case, the limit of all data, we have the following conjecture:

**Conjecture 1** *In case (iii), the limit of all data samples $\nu \to 0$, the optimal noise distribution is such that it is all concentrated at the set of those $\boldsymbol{\xi}$ which are given by*

$$\begin{aligned} &\arg\max_{\boldsymbol{\xi}} p_d(\boldsymbol{\xi}) \mathrm{tr}\left( (\boldsymbol{g}(\xi)\boldsymbol{g}(\xi)^\top)^{-1} \right)^{-1} \\ &\qquad = p_d(\boldsymbol{\xi}) \|\boldsymbol{g}(\xi)\|^2 \quad \textit{for a scalar parameter} \end{aligned} \tag{11}$$

Informal and heuristic *"proof"*:

We have the $\mathrm{MSE}_{\mathrm{NCE}}(T, \nu, p_n) = \frac{\nu+1}{T}\mathrm{tr}(\boldsymbol{I}^{-1} - \frac{\nu+1}{\nu}(\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1}))$.

Given the term up until $\nu^{-1}$ in the MSE, we will use Taylor expansions up to order 2 throughout the proof, in anticipation that the MSE will be expanded until order 1.

Note that in this no noise limit, the assumption made by Gutmann and Hyvärinen (2012) that $p_n$ is non-zero whenever $p_d$ is nonzero is not true for this optimal $p_n$, which reduces the rigour of this analysis. (This we denote by heuristic approximation 1.)

- Taylor expansion of the discriminator

$$1 - D(\boldsymbol{x}) = \frac{\nu p_n(\boldsymbol{x})}{p_d(\boldsymbol{x}) + \nu p_n(\boldsymbol{x})} = \frac{1}{1 + \frac{1}{\nu}\frac{p_d}{p_n}(\boldsymbol{x})} = \nu \frac{p_n}{p_d}(\boldsymbol{x}) - \nu^2 \frac{p_n^2}{p_d^2}(\boldsymbol{x}) + \circ(\nu^2)$$

- Evaluating the integrals $\boldsymbol{m}$, $\boldsymbol{I}$

$$\begin{aligned} \boldsymbol{m} &= \int \boldsymbol{g}(\boldsymbol{x}) p_d(\boldsymbol{x}) \Big(1 - D(\boldsymbol{x})\Big) d\boldsymbol{x} = \int \boldsymbol{g}(\boldsymbol{x}) p_d(\boldsymbol{x}) \left( \nu \frac{p_n}{p_d}(\boldsymbol{x}) - \nu^2 \frac{p_n^2}{p_d^2}(\boldsymbol{x}) + \circ(\nu^2) \right) d\boldsymbol{x} \\ &= \nu \boldsymbol{m}_n - \nu^2 \boldsymbol{b} + \circ(\nu^2) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{m}_n &= \int \boldsymbol{g}(\boldsymbol{x}) p_n(\boldsymbol{x}) d\boldsymbol{x} = 0 \\ \boldsymbol{b} &= \int \boldsymbol{g}(\boldsymbol{x}) \frac{p_n^2}{p_d}(\boldsymbol{x}) d\boldsymbol{x} \ . \end{aligned}$$

Similarly,

$$\begin{aligned} \boldsymbol{I} &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p_d(\boldsymbol{x}) \Big(1 - D(\boldsymbol{x})\Big) d\boldsymbol{x} = \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p_d(\boldsymbol{x}) \left( \nu \frac{p_n}{p_d}(\boldsymbol{x}) - \nu^2 \frac{p_n^2}{p_d^2}(\boldsymbol{x}) + \circ(\nu^2) \right) d\boldsymbol{x} \\ &= \nu \boldsymbol{I}_n - \nu^2 \boldsymbol{B} + \circ(\nu^2) \end{aligned}$$

where the Fisher-score covariance (Fisher information) is $\boldsymbol{I}_F$ and we use shorthand notations $A$ and $B$ for the remaining integrals:

$$\begin{aligned} \boldsymbol{I}_n &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top p_n(\boldsymbol{x}) d\boldsymbol{x} \\ \boldsymbol{B} &= \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top \frac{p_n^2}{p_d}(\boldsymbol{x}) d\boldsymbol{x} \ . \end{aligned}$$

- Taylor expansion of $\boldsymbol{I}^{-1}$

$$
\begin{aligned}
\boldsymbol{I}^{-1} &= \Big(\nu \boldsymbol{I}_n - \nu^2 \boldsymbol{B} + \circ(\nu^2)\Big)^{-1} \\
&= \Big(\nu \boldsymbol{I}_n (\mathbf{Id} - \nu \boldsymbol{I}_n^{-1} \boldsymbol{B}) + \circ(\nu^2)\Big)^{-1} \\
&= \nu^{-1} \boldsymbol{I}_n^{-1} \Big(\mathbf{Id} + \nu \boldsymbol{I}_n^{-1} \boldsymbol{B} + \nu^2 (\boldsymbol{I}_n^{-1} \boldsymbol{B})^2 + \nu^3 (\boldsymbol{I}_n^{-1} \boldsymbol{B})^3 + \circ(\nu^3)\Big) + \circ(\nu^2) \\
&= \nu^{-1} \boldsymbol{I}_n^{-1} + \nu^0 \boldsymbol{I}_n^{-2} \boldsymbol{B} + \nu^1 \boldsymbol{I}_n^{-1} (\boldsymbol{I}_n^{-2} \boldsymbol{B})^2 + \nu^2 \boldsymbol{I}_n^{-1} (\boldsymbol{I}_n^{-2} \boldsymbol{B})^3 + \circ(\nu^2)
\end{aligned}
$$

- Evaluating the $\mathrm{MSE}_{\mathrm{NCE}}$

$$
\begin{aligned}
&\boldsymbol{I}^{-1} \boldsymbol{m} \boldsymbol{m}^\top \boldsymbol{I}^{-1} = \\
&\nu^0 (\boldsymbol{I}_n^{-1} \boldsymbol{m}_n \boldsymbol{m}_n^T \boldsymbol{I}_n^{-1}) + \nu^2 (\boldsymbol{I}_n^{-1} \boldsymbol{b}_n \boldsymbol{b}_n^T \boldsymbol{I}_n^{-1} + \boldsymbol{I}_n^{-2} \boldsymbol{B} \boldsymbol{b}_n \boldsymbol{b}_n^T \boldsymbol{I}_n^{-2} \boldsymbol{B}) + \circ(\nu^2)
\end{aligned}
$$

by plugging in the Taylor expansions of $\boldsymbol{I}^{-1}$ and $\boldsymbol{m}$ and retaining only terms up to the second order. Hence, the second term of the MSE without the trace is

$$
\begin{aligned}
&\Big(\nu^1 T^{-1} + \nu^0 2T^{-1} + \nu^{-1} T^{-1}\Big) \boldsymbol{I}^{-1} \boldsymbol{m} \boldsymbol{m}^\top \boldsymbol{I}^{-1} \\
&= \Big(\nu^1 T^{-1} + \nu^0 2T^{-1} + \nu^{-1} T^{-1}\Big) \Big(\nu^0 (\boldsymbol{I}_n^{-1} \boldsymbol{m}_n \boldsymbol{m}_n^T \boldsymbol{I}_n^{-1}) + \nu^2 (\boldsymbol{I}_n^{-1} \boldsymbol{b}_n \boldsymbol{b}_n^T \boldsymbol{I}_n^{-1} + \boldsymbol{I}_n^{-2} \boldsymbol{B} \boldsymbol{b}_n \boldsymbol{b}_n^T \boldsymbol{I}_n^{-2} \boldsymbol{B}) + \circ(\nu^2)\Big) \\
&= \nu^{-1} \frac{1}{T} (\boldsymbol{I}_n^{-1} \boldsymbol{m}_n \boldsymbol{m}_n^T \boldsymbol{I}_n^{-1}) + \nu^0 \frac{1}{T} (2\boldsymbol{I}_n^{-1} \boldsymbol{m}_n \boldsymbol{m}_n^T \boldsymbol{I}_n^{-1}) + \nu^1 \frac{1}{T} (\boldsymbol{I}_n^{-1} \boldsymbol{b}_n \boldsymbol{b}_n^T \boldsymbol{I}_n^{-1} + \boldsymbol{I}_n^{-2} \boldsymbol{B} \boldsymbol{b}_n \boldsymbol{b}_n^T \boldsymbol{I}_n^{-2} \boldsymbol{B} + \boldsymbol{I}_n^{-1} \boldsymbol{m}_n \boldsymbol{m}_n^T \boldsymbol{I}_n^{-1}) + \circ(\nu)
\end{aligned}
$$

and the first term of the MSE without the trace is

$$
\begin{aligned}
&\Big(\nu^0 T^{-1} + \nu^1 T^{-1}\Big) \mathrm{tr}(\boldsymbol{I}^{-1}) \\
&= \Big(\nu^0 T^{-1} + \nu^1 T^{-1}\Big) \Big(\nu^{-1} \boldsymbol{I}_n^{-1} + \nu^0 \boldsymbol{I}_n^{-2} \boldsymbol{B} + \nu^1 \boldsymbol{I}_n^{-1} (\boldsymbol{I}_n^{-2} \boldsymbol{B})^2 + \nu^2 \boldsymbol{I}_n^{-1} (\boldsymbol{I}_n^{-2} \boldsymbol{B})^3 + \circ(\nu^2)\Big) \\
&= \nu^{-1} \frac{1}{T} \boldsymbol{I}_n^{-1} + \nu^0 \frac{1}{T} (\boldsymbol{I}_n^{-2} \boldsymbol{B} + \boldsymbol{I}_n^{-1}) + \nu^1 \frac{1}{T} [\boldsymbol{I}_n^{-1} (\boldsymbol{I}_n^{-1} \boldsymbol{B})^2 + \boldsymbol{I}_n^{-2} \boldsymbol{B}] + \circ(\nu) \ .
\end{aligned}
$$

Subtracting the second term from the first term and applying the trace, we finally write the MSE:

$$
\begin{aligned}
\mathrm{MSE}_{\mathrm{NCE}} = \mathrm{tr}(&\nu^{-1} \frac{1}{T} (\boldsymbol{I}_n^{-1} - \boldsymbol{I}_n^{-1} \boldsymbol{m}_n \boldsymbol{m}_n^T \boldsymbol{I}_n^{-1}) + \nu^0 \frac{1}{T} (\boldsymbol{I}_n^{-2} \boldsymbol{B} + \boldsymbol{I}_n^{-1} - 2\boldsymbol{I}_n^{-1} \boldsymbol{m}_n \boldsymbol{m}_n^T \boldsymbol{I}_n^{-1}) + \\
&\nu^1 \frac{1}{T} [\boldsymbol{I}_n^{-1} (\boldsymbol{I}_n^{-1} \boldsymbol{B})^2 + \boldsymbol{I}_n^{-2} \boldsymbol{B} - \boldsymbol{I}_n^{-1} \boldsymbol{b}_n \boldsymbol{b}_n^T \boldsymbol{I}_n^{-1} - \boldsymbol{I}_n^{-2} \boldsymbol{B} \boldsymbol{b}_n \boldsymbol{b}_n^T \boldsymbol{I}_n^{-2} \boldsymbol{B} - \boldsymbol{I}_n^{-1} \boldsymbol{m}_n \boldsymbol{m}_n^T \boldsymbol{I}_n^{-1}] + \circ(\nu)) \ .
\end{aligned}
$$

Rewriting $\boldsymbol{I}_n^{-1} = \boldsymbol{I}_n^{-1} \boldsymbol{I}_n \boldsymbol{I}_n^{-1}$, using the circular invariance of the trace operator and stopping at order $\nu^0$, we get:

$$
\begin{aligned}
\mathrm{MSE}_{\mathrm{NCE}} &= \nu^{-1} \frac{1}{T} \langle \boldsymbol{I}_n^{-2}, \boldsymbol{I}_n - \boldsymbol{m}_n \boldsymbol{m}_n^T \rangle + \nu^0 \frac{1}{T} \langle \boldsymbol{I}_n^{-2}, \boldsymbol{B} + \boldsymbol{I}_n - 2\boldsymbol{m}_n \boldsymbol{m}_n^T \rangle + \circ(1) \\
&= \nu^{-1} \frac{1}{T} \langle \boldsymbol{I}_n^{-2}, \mathrm{Var}_{N \sim p_n} \boldsymbol{g}(\mathbf{N}) \rangle + \nu^0 \frac{1}{T} \langle \boldsymbol{I}_n^{-2}, \boldsymbol{B} + \boldsymbol{I}_n - 2\boldsymbol{m}_n \boldsymbol{m}_n^T \rangle + \circ(1) \ . \qquad (12)
\end{aligned}
$$

- Optimize the $\mathrm{MSE}_{\mathrm{NCE}}$ w.r.t. $p_n$

  Looking at the above MSE, the dominant term of order $\nu^{-1}$ is $\langle \boldsymbol{I}_n^{-2}, \mathrm{Var}_{N\sim p_n}\boldsymbol{g}(\mathbf{N})\rangle \geq 0$ is minimized when it is 0, that is, when $\boldsymbol{g}$ is constant in the support of $p_n$. Typically this means that $p_n$ is concentrated on a set of zero measure. In the 1D case, such case is typically the Dirac delta $p_n = \delta_z$, or a distribution with two deltas in case of symmetrical $\boldsymbol{g}$.

  We can plug this in the terms of the next order $\nu^0$, which remain to be minimized:

$$\begin{aligned}\langle \boldsymbol{I}_n^{-2}, \boldsymbol{B} + \boldsymbol{I}_n - 2\boldsymbol{m}_n\boldsymbol{m}_n^T\rangle &= \langle \boldsymbol{I}_n^{-2}, \boldsymbol{B} - \boldsymbol{I}_n + 2\boldsymbol{I}_n - 2\boldsymbol{m}_n\boldsymbol{m}_n^T\rangle \\ &= \langle \boldsymbol{I}_n^{-2}, \boldsymbol{B} - \boldsymbol{I}_n + 2\mathrm{Var}_{N\sim p_n}\boldsymbol{g}(\mathbf{N})\rangle \\ &= \langle \boldsymbol{I}_n^{-2}, \boldsymbol{B} - \boldsymbol{I}_n\rangle\end{aligned}$$

  given we chose $p_n$ so that the variance is 0.

  The integrands of $\boldsymbol{B}$ and $\boldsymbol{I}$ respectively involve $p_n^2$ and $p_n$. Because $p_n$ is concentrated on a set of zero measure (Dirac-like), the term in $\boldsymbol{B}$ dominates the term in $\boldsymbol{I}$. This is because if we consider the $p_n$ as the limit of a sequence of some proper pdf's, the value of the pdf gets infinite in the support of that pdf in the limit, and thus $p_n^2$ is infinitely larger than $p_n$. Hence we are left with $\langle \boldsymbol{I}_n^{-2}, \boldsymbol{B}\rangle$.

  The integral with respect to $p_n$ simplifies to simply evaluating the $\boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top/p_d(\boldsymbol{x})$ the support of $p_n$. Since we know that $\boldsymbol{g}(\boldsymbol{x})$ is constant in that set, the main question is whether $p_d$ is constant in that set as well. Here, we heuristically assume that it is; this is intuitively appealing in many cases, if not necessarily true. (This we denote by heuristic approximation 2.)

  Thus, we have

$$\int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top \frac{\delta_z^2}{p_d}(\boldsymbol{x})d\boldsymbol{x} \approx c\, \boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top \frac{1}{p_d(\boldsymbol{z})}$$

  for some constant $c$ taking into account the effect of squaring of $p_d$ (it is ultimately infinite, but the reasoning is still valid in any sequence going to the limit.)

  Next we make heuristic approximation 3: we neglect any problems of inversion of singular, rank 1 matrices (note this is not a problem in the 1D case), and obtain further

$$\langle \boldsymbol{I}_n^{-2}, \boldsymbol{B}\rangle \approx \mathrm{tr}\left((\boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top)^{-1}\boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top \frac{1}{p_d(\boldsymbol{z})}(\boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top)^{-1}\right) \approx \frac{1}{p_d(\boldsymbol{z})}\mathrm{tr}\left((\boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top)^{-1}\right) \tag{13}$$

  Minimizing this term is equivalent to the following maximization setup (still applying heuristic approximation 3):

$$\arg\max_{\boldsymbol{\xi}} p_d(\boldsymbol{\xi})\mathrm{tr}\left((\boldsymbol{g}(\xi)\boldsymbol{g}(\xi)^\top)^{-1}\right)^{-1}$$

  Those points $z$ obtained by the above condition are the best candidates for $p_n$ to concentrate its mass on.

  We arrived this result by making three heuristic approximations as explained above; we hope to be able to remove some of them in future work.

  *Addendum:* In fact, the informal proof just given implies a stronger result than announced in the Conjecture of the main paper. Inside the set given by the Conjecture (and the arg max above), the probability mass should further be concentrated to a set where $\boldsymbol{g}(\boldsymbol{x})$ is constant, as was shown to be necessary earlier in the proof. This in fact explains the result on top row of Fig 2b, where the probability mass has to all go to one of the optimal points since $\boldsymbol{g}(\boldsymbol{x})$ takes different values in them (see Section 2 of supplementary material for the functional forms of the score functions). (In contrast, in Fig 2a, $\boldsymbol{g}(\boldsymbol{x})$ takes equal values in the two optimal points and therefore the mass is distributed to them equally.)

## 5 Optimal Noise for Estimating a Distribution: Proofs

So far, we have optimized hyperparameters (such as the noise distribution) so that the reduce the uncertainty of the *parameter* estimation, measured by the Mean Squared Error $\mathbb{E}\big[\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^2\big] = \frac{1}{T_d}\mathrm{tr}(\boldsymbol{\Sigma})$.

Sometimes, we might wish to reduce the uncertainty of the *distribution* estimation, which we can measure using the Kullback-Leibler (KL) divergence $\mathbb{E}\big[\mathcal{D}_{\mathrm{KL}}(p_d, p_{\hat{\boldsymbol{\theta}}_T})\big]$.

We can specify this error, by using the Taylor expansion of the estimated $\hat{\theta}_T$ near optimality, given in **?**:

$$\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^* = \boldsymbol{z} + O(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^2) \tag{14}$$

where $z \sim \mathcal{N}(0, \frac{1}{T_d}\boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is the asymptotic variance matrix.

We can similarly take the Taylor expansion of the KL divergence with respect to its second argument, near optimality:

$$\begin{aligned}
J(\hat{\theta}_T) &:= \mathcal{D}_{\mathrm{KL}}(p_d, p_{\hat{\boldsymbol{\theta}}_T}) \\
&= J(\boldsymbol{\theta}^*) + < \nabla_\theta J(\theta^*), \hat{\theta}_T - \theta^* > + \frac{1}{2} < (\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*), \nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^*)\,(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) > + O(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^3) \\
&= J(\boldsymbol{\theta}^*) + < \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^*), \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) > + \frac{1}{2}\big\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\big\|^2_{\nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^*)} + O(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^3)
\end{aligned}$$

Note that some simplifications occur:

- $J(\boldsymbol{\theta}^*) = \mathcal{D}_{\mathrm{KL}}(p_{\boldsymbol{\theta}^*}, p_{\boldsymbol{\theta}^*}) = 0$
- $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^*) = 0$ as the gradient the KL divergence at $\boldsymbol{\theta}^*$ is the mean of the (negative) Fisher score, which is null.
- $\nabla^2_{\boldsymbol{\theta}} J(\theta^*) = \boldsymbol{I}_F$

Plugging in the estimation error 14 into the distribution error yields:

$$\begin{aligned}
J(\hat{\boldsymbol{\theta}}_T) &= \frac{1}{2}\Big\|\boldsymbol{z} + O(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^2)\Big\|^2_{\boldsymbol{I}_F} + O(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^3) \\
&= \frac{1}{2}\Big(\|\boldsymbol{z}\|^2_{\boldsymbol{I}_F} + 2 < \boldsymbol{z}, O(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2) >_{\boldsymbol{I}_F} + \big\|O(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2)\big\|^2_{\boldsymbol{I}_F}\Big) + O(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^3) \\
&= \frac{1}{2}\|\boldsymbol{z}\|^2_{\boldsymbol{I}_F} + O(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^2)
\end{aligned}$$

by truncating the Taylor expansion to the first order. Hence up to the first order, the expectation yields:

$$\begin{aligned}
\mathbb{E}\big[\mathcal{D}_{\mathrm{KL}}(p_d, p_{\hat{\boldsymbol{\theta}}_T})\big] &= \frac{1}{2}\mathbb{E}\big[\|\boldsymbol{z}\|^2_{\boldsymbol{I}_F}\big] = \frac{1}{2}\mathbb{E}\big[\boldsymbol{z}^T\boldsymbol{I}_F\boldsymbol{z}\big] = \frac{1}{2}\mathbb{E}\big[\mathrm{tr}(\boldsymbol{z}^T\boldsymbol{I}_F\boldsymbol{z})\big] = \frac{1}{2}\mathbb{E}\big[\mathrm{tr}(\boldsymbol{I}_F\boldsymbol{z}\boldsymbol{z}^T)\big] \\
&= \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_F\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]) = \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_F\mathrm{Var}[\boldsymbol{z}]) = \frac{1}{2T_d}\mathrm{tr}(\boldsymbol{I}_F\boldsymbol{\Sigma})
\end{aligned}$$

Finding the optimal noise that minimizes the distribution error means minimizing $\frac{1}{T_d}\mathrm{tr}(\Sigma I_F)$. Contrast that with the optimal noise that minimizes the parameter estimation error (asymptotic variance) $\frac{1}{T_d}\mathrm{tr}(\Sigma)$. We can reprise each of the three limit cases from the previous proofs, and derive novel optimal noise distributions:

**Theorem 2** *In cases:*
*(i) when the noise distribution is a (infinitesimal) perturbation of the data distribution* $\frac{p_d}{p_n} = 1 + \epsilon(x)$*;*
*(ii) in the limit of all noise samples* $\nu \to \infty$*;*
*the optimal noise distributions match. They are given by:*

$$p_n^{\text{opt}}(\boldsymbol{x}) \propto p_d(\boldsymbol{x}) \| \boldsymbol{I}_F^{-\frac{1}{2}} \boldsymbol{g}(\boldsymbol{x}) \| \ . \tag{15}$$

*Proof:* case of $\nu \to \infty$

We recall the asymptotic variance $\frac{1}{T_d}\Sigma$ in the all-noise limit is given by equation 5 at the first order and without the trace. Multiplying by $I_F$ introduces no additional dependency in $p_n$, hence we retain the only term dependent that was dependent on $p_n$, $\boldsymbol{I}_F^{-2} \int \boldsymbol{g}(x)\boldsymbol{g}(x)^\top \frac{p_d^2}{p_n}(\boldsymbol{x})d\boldsymbol{x}$, multiply it with $I_F$ and take the trace. This yields the following cost to minimize:

$$J(p_n) = \frac{1}{T} \int \| I_F^{-\frac{1}{2}} \boldsymbol{g}(\boldsymbol{x}) \|^2 \frac{p_d^2}{p_n}(\boldsymbol{x}) d\boldsymbol{x} \tag{16}$$

with respect to $p_n$. As in previous proofs, we compute the variational (Fréchet) derivative together with the Lagrangian of the constraint $\int p_n(\boldsymbol{x}) = 1$ (with $\lambda$ denoting the Lagrangian multiplier) to obtain

$$\delta_{p_n} J = -\| \boldsymbol{I}_F^{-\frac{1}{2}} \boldsymbol{g} \|^2 \frac{p_d^2}{p_n^2} + \lambda \ . \tag{17}$$

Setting this to zero and taking into account the non-negativity of $p_n$ gives

$$p_n(\boldsymbol{x}) = \| \boldsymbol{I}_F^{-\frac{1}{2}} \boldsymbol{g}(x) \| p_d(\boldsymbol{x}) / Z \tag{18}$$

where $Z = \int \| \boldsymbol{I}_F^{-\frac{1}{2}} \boldsymbol{g}(x) \| p_d(\boldsymbol{x}) d\boldsymbol{x}$ is the normalization constant. This is thus the optimal noise distribution, as a first-order approximation.

**Conjecture 2** *In case (iii), the limit of all data samples* $\nu \to 0$*, the optimal noise distribution is such that it is all concentrated at the set of those* $\boldsymbol{\xi}$ *which are given by*

$$\begin{aligned} &\arg\max_{\boldsymbol{\xi}} p_d(\boldsymbol{\xi}) \text{tr}\left( (\boldsymbol{g}(\xi)\boldsymbol{g}(\xi)^\top)^{-\frac{1}{2}} \right)^{-1} \\ &= p_d(\boldsymbol{\xi}) \| \boldsymbol{g}(\xi) \| \quad \textit{for a scalar parameter} \end{aligned} \tag{19}$$

*Proof:* case of $\nu \to$

By the same considerations, we can obtain the optimal noise that minimizes the asymptotic error in distribution space in the all-data limit, using equation with a multiplication by $I_F$ inside the the trace. This leads to the result.

## 6 Numerical Validation of the Predicted Distribution Error

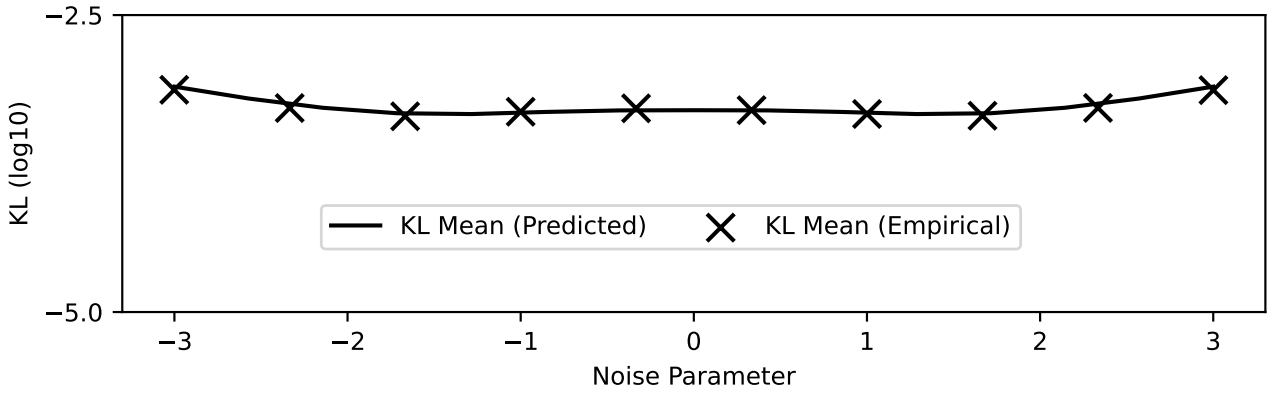



Figure 2: KL vs. the noise parameter (Gaussian Mean). The noise proportion is fixed at 50%.

We here numerically validate our formulae predicting the asymptotic estimation error in distribution space $\mathcal{D}_{\mathrm{KL}}(p_d, p_{\hat{\theta}_{\mathrm{NCE}}})$, when the noise is constrained within a parametric family containing the data; here, the model is a one-dimensional centered Gaussian with unit variance, parameterized by its mean.