The Optimal Noise in Noise-Contrastive Learning Is Not What You Think (Supplementary Material)

Omar Chehab[1] **Alexandre Gramfort**[1] **Aapo Hyvärinen**[2]

[1]Université Paris-Saclay, Inria, CEA, Palaiseau, France
[2]Department of Computer Science, University of Helsinki, Finland

1 VISUALIZATIONS OF THE MSE LANDSCAPE

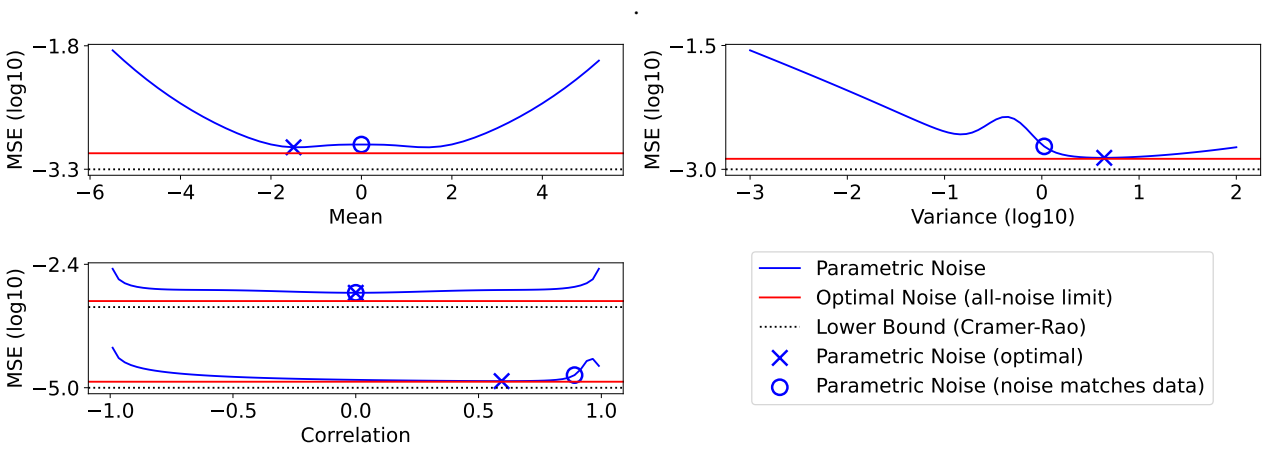


Figure 1: MSE vs. the noise parameter. Top left panel for model (i), Gaussian mean; Top right panel for model (ii), Gaussian variance; Bottom left for model (iii), Gaussian correlation.

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.

Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022), PMLR 180:1–16.

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:

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

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 β and π (noise distribution) or ν (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

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

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

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

The objective function thus reduces to

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

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

Note that in that case where $p_n = p_d$, we can compare the MSE achieved by NCE (using T_d data samples and T_n noise samples) with the MSE achieved my MLE (using T_d data samples):

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

which is known from [Gutmann and Hyvärinen, 2012, Pihlaja et al., 2010].

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 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

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

Proof: case where $\nu \to \infty$.

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:

$$\begin{aligned} \text{MSE}_{\text{NCE}}(T, \gamma, p_n) &= \frac{\gamma+1}{\gamma T}\text{tr}(\boldsymbol{I}^{-1}) - \frac{(\gamma+1)^2}{T\gamma}\text{tr}(\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1}) &(3)\\ &= \left(\gamma^{-1}T^{-1} + \gamma^0 T^{-1}\right)\text{tr}(\boldsymbol{I}^{-1}) - \left(\gamma^{-1}T^{-1} + \gamma^0 2T^{-1} + \gamma^1 T^{-1}\right)\text{tr}(\boldsymbol{I}^{-1}\boldsymbol{m}\boldsymbol{m}^\top\boldsymbol{I}^{-1}) &(4)\end{aligned}$$

Given the term up until γ^{-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 $\boldsymbol{m}$, $\boldsymbol{I}$

$$\begin{aligned}\boldsymbol{m} &= \int \boldsymbol{g}(\boldsymbol{x})p_d(\boldsymbol{x})\Big(1 - D(\boldsymbol{x})\Big)d\boldsymbol{x} = \int \boldsymbol{g}(\boldsymbol{x})p_d(\boldsymbol{x})\left(1 - \gamma\frac{p_d}{p_n}(\boldsymbol{x}) + \gamma^2\frac{p_d^2}{p_n^2}(\boldsymbol{x}) + \circ(\gamma^2)\right)d\boldsymbol{x}\\ &= \boldsymbol{m}_F - \gamma\boldsymbol{a} + \gamma^2\boldsymbol{b} + \circ(\gamma^2) &(5)\end{aligned}$$

where $\boldsymbol{m}_F$ is the Fisher-score mean of the (possibly unnormalized) model and we use shorthand notations a and b for the remaining integrals:

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

Similarly,

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

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

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

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

$$\begin{aligned}\boldsymbol{I}^{-1} &= \Big(\boldsymbol{I}_F - \gamma\boldsymbol{A} + \gamma^2\boldsymbol{B} + \circ(\gamma^2)\Big)^{-1} \\ &= \Big(\boldsymbol{I}_F(\mathbf{Id} - \gamma\boldsymbol{I}_F^{-1}\boldsymbol{A} + \gamma^2\boldsymbol{I}_F^{-1}\boldsymbol{B}) + \circ(\gamma^2)\Big)^{-1} \\ &= \boldsymbol{I}_F^{-1}\Big(\mathbf{Id} - \gamma\boldsymbol{I}_F^{-1}\boldsymbol{A} + \gamma^2\boldsymbol{I}_F^{-1}\boldsymbol{B}\Big)^{-1} + \circ(\gamma^2) \\ &= \boldsymbol{I}_F^{-1}\Big(\mathbf{Id} + \gamma\boldsymbol{I}_F^{-1}\boldsymbol{A} + \gamma^2((\boldsymbol{I}_F^{-1}\boldsymbol{A})^2 - \boldsymbol{I}_F^{-1}\boldsymbol{B}) + \circ(\gamma^2)\Big) + \circ(\gamma^2) \\ &= \boldsymbol{I}_F^{-1} + \gamma\boldsymbol{I}_F^{-2}\boldsymbol{A} + \gamma^2(\boldsymbol{I}_F^{-1}(\boldsymbol{I}_F^{-1}\boldsymbol{A})^2 - \boldsymbol{I}_F^{-2}\boldsymbol{B}) + \circ(\gamma^2) \qquad (6)\end{aligned}$$

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

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

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

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

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:

$$\mathrm{MSE}_{\mathrm{NCE}} = \mathrm{tr}\Big(\gamma^{-1}\frac{1}{T}\big(\boldsymbol{I}_F^{-1} - \boldsymbol{I}_F^{-1}\boldsymbol{m}_F\boldsymbol{m}_F^\top\boldsymbol{I}_F^{-1}\big) + \gamma^0\frac{1}{T}\big(\boldsymbol{I}_F^{-2}\boldsymbol{A} + \boldsymbol{I}_F^{-1} - 2\boldsymbol{I}_F^{-1}\boldsymbol{m}_F\boldsymbol{m}_F^\top\boldsymbol{I}_F^{-1}\big)\Big) + \circ(\gamma) \qquad (7)$$

- 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

$$J(p_n) = \frac{1}{T}\int \|I_F^{-1}\boldsymbol{g}(\boldsymbol{x})\|^2\frac{p_d^2}{p_n}(\boldsymbol{x})d\boldsymbol{x} \qquad (8)$$

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 λ denoting the Lagrangian multiplier) to obtain

$$\delta_{p_n} J = -\|\boldsymbol{I}_F^{-1}\boldsymbol{g}\|^2 \frac{p_d^2}{p_n^2} + \lambda \ . \tag{9}$$

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

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

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.

- Compute the MSE gaps at optimality
  Plugging this optimal p_n into the formula of $\mathrm{MSE}_{\mathrm{NCE}}$ and subtracting the Cramer-Rao MSE (which is a lower bound for a normalized model), 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^{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 ϵ would be a constant.

- Taylor expansion of the discriminator

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

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

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

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

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

  Similarly,

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

  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 $\mathrm{MSE}_{\mathrm{NCE}}$

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

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:

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

- 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

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

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

$$J(p_n) = \frac{1}{T}\int \|I_F^{-1}\boldsymbol{g}(\boldsymbol{x})\|^2\frac{p_d^2}{p_n}(\boldsymbol{x})d\boldsymbol{x} \tag{11}$$

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

$$p_n(\boldsymbol{x}) = \|\boldsymbol{I}_F^{-1}\boldsymbol{g}(x)\|p_d(\boldsymbol{x})/Z \tag{12}$$

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*

$$\begin{aligned} &\arg\max_{\boldsymbol{\xi}} p_d(\boldsymbol{\xi}) \mathrm{tr}\Big((\boldsymbol{g}(\xi)\boldsymbol{g}(\xi)^\top)^{-1} \Big)^{-1} \\ &\text{s.t.} \quad \boldsymbol{g}(\xi) = \text{constant} \end{aligned} \tag{13}$$

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 ν^{-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 $\boldsymbol{m}$, $\boldsymbol{I}$

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

  where

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

  Similarly,

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

  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 $\mathrm{MSE}_{\mathrm{NCE}}$

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

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

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

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:

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

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 ν^0, we get:

$$
\begin{aligned}
\mathrm{MSE}_{\mathrm{NCE}} &= \nu^{-1} \frac{1}{T} \langle \boldsymbol{I}_n^{-2}, \boldsymbol{I}_n - \boldsymbol{m}_n \boldsymbol{m}_n^\top \rangle + \nu^0 \frac{1}{T} \langle \boldsymbol{I}_n^{-2}, \boldsymbol{B} + \boldsymbol{I}_n - 2 \boldsymbol{m}_n \boldsymbol{m}_n^\top \rangle + \circ(1) \\
&= \nu^{-1} \frac{1}{T} \langle \boldsymbol{I}_n^{-2}, \mathrm{Var}_{N \sim p_n} \boldsymbol{g}(\mathbf{N}) \rangle + \nu^0 \frac{1}{T} \langle \boldsymbol{I}_n^{-2}, \boldsymbol{B} + \boldsymbol{I}_n - 2 \boldsymbol{m}_n \boldsymbol{m}_n^\top \rangle + \circ(1) \ . \qquad (14)
\end{aligned}
$$

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

  Looking at the above MSE, the dominant term of order ν^{-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 ν^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_n (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 further obtain

$$\langle \boldsymbol{I}_n^{-2}, \boldsymbol{B}\rangle \approx \mathrm{tr}\left((\boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top)^{-1}\boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top \frac{1}{p_d(\boldsymbol{z})}(\boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top)^{-1}\right) \approx \frac{1}{p_d(\boldsymbol{z})}\mathrm{tr}\left((\boldsymbol{g}(\boldsymbol{z})\boldsymbol{g}(\boldsymbol{z})^\top)^{-1}\right) . \tag{15}$$

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.

Numerically, evaluating the optimal noise in the all-data limit requires computing a weight $w(\boldsymbol{x}) = \mathrm{tr}\left((\boldsymbol{g}(\xi)\boldsymbol{g}(\xi)^\top)^{-1}\right)^{-1}$ that is intractable in dimensions bigger than 1, due to the singularity of the rank 1 matrix. We can avoid this numerically by introducing an (infinitesimal) perturbation $\epsilon > 0$ which removes the singularity problem:

$$\begin{aligned}w_\epsilon(\xi) &= \mathrm{tr}\left((\boldsymbol{g}(\xi)\boldsymbol{g}(\xi)^\top + \epsilon\mathrm{Id})^{-1}\right)^{-1} \\ &= \mathrm{tr}\left(\epsilon^{-1}\mathrm{Id} - \frac{1}{\epsilon^2 + \epsilon\boldsymbol{g}(\xi)^\top\mathrm{Id}\boldsymbol{g}(\xi)}\boldsymbol{g}(\xi)\boldsymbol{g}(\xi)^\top\right)^{-1} \quad \text{by the Sherman-Morrison formula} \\ &= \left(\epsilon^{-1}d - \frac{1}{\epsilon^2 + \epsilon\|\boldsymbol{g}(\xi)\|^2}\|\boldsymbol{g}(\xi)\|^2\right)^{-1} \\ &= \left(\epsilon^{-1}(d-1) + \epsilon^0\frac{1}{\|\boldsymbol{g}(\xi)\|^2} + \epsilon^1\frac{-1}{\|\boldsymbol{g}(\xi)\|^4} + O(\epsilon^2)\right)^{-1} \quad \text{by Taylor expansion} \\ &= \epsilon\frac{1}{d-1} + \epsilon^2\frac{-1}{\|\boldsymbol{g}(\xi)\|^2(d-1)^2} + \epsilon^3\frac{(2-d)}{\|\boldsymbol{g}(\xi)\|^4(d-1)^3} + O(\epsilon^4) \quad \text{by further Taylor expansion}\end{aligned}$$

where we go up to order 3 to ensure the weight $w_\epsilon(\xi)$ is positive. Finally, we can approximate the $\arg\max$ operator with its relaxation $\text{soft}\arg\max^\epsilon(x) = \frac{e^{\frac{x}{\epsilon}}}{\int e^{\frac{x}{\epsilon}} dx}$, so that

$$p_n(\boldsymbol{x}) \approx \text{soft}\arg\max^{\epsilon_1} \left(p_d(\boldsymbol{x}) w_{\epsilon_2}(\boldsymbol{x})\right)$$

where $(\epsilon_1, \epsilon_2) \in (\mathbb{R}_+^*)^2$ are two hyperparameters taken close to zero.

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 Gutmann and Hyvärinen [2012]:

$$\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^* = \boldsymbol{z} + O(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^2) \tag{16}$$

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:

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

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 16 into the distribution error yields:

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

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

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

Note that this is a general and known result which is applicable beyond the KL divergence: for any divergence, the 0th order term is null as it measures the divergence between the data distribution and itself, the 1st order term is null in expectation if the estimator $\hat{\theta}_T$ is asymptotically unbiased, which leaves an expected error given by the 2nd-order term $\frac{1}{2T_d}\mathrm{tr}(\nabla^2 J\boldsymbol{\Sigma})$ where J is the chosen divergence. Essentially, one would replace the Fisher Information above, which is the Hessian for a forward-KL divergence, by the Hessian for a given divergence.

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

$$p_n^{\text{opt}}(\boldsymbol{x}) \propto p_d(\boldsymbol{x}) \| \boldsymbol{I}_F^{-\frac{1}{2}} \boldsymbol{g}(\boldsymbol{x}) \| \ . \tag{17}$$

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 7 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:

$$J(p_n) = \frac{1}{T} \int \| I_F^{-\frac{1}{2}} \boldsymbol{g}(\boldsymbol{x}) \|^2 \frac{p_d^2}{p_n}(\boldsymbol{x}) d\boldsymbol{x} \tag{18}$$

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 λ denoting the Lagrangian multiplier) to obtain

$$\delta_{p_n} J = -\| \boldsymbol{I}_F^{-\frac{1}{2}} \boldsymbol{g} \|^2 \frac{p_d^2}{p_n^2} + \lambda \ . \tag{19}$$

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

$$p_n(\boldsymbol{x}) = \| \boldsymbol{I}_F^{-\frac{1}{2}} \boldsymbol{g}(x) \| p_d(\boldsymbol{x}) / Z \tag{20}$$

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.

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

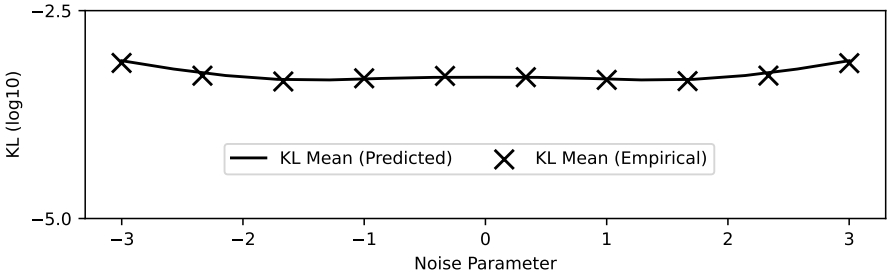
Conjecture 2 *In the limit of Conjecture 1 the noise distribution minimizing the expected Kullback-Leibler divergence is such that it is all concentrated at the set of those $\boldsymbol{\xi}$ which are given by*

$$\begin{aligned} &\arg\max_{\boldsymbol{\xi}} p_d(\boldsymbol{\xi}) \text{tr} \left((\boldsymbol{g}(\xi)\boldsymbol{g}(\xi)^\top)^{-\frac{1}{2}} \right)^{-1} \\ &\text{s.t.} \quad \boldsymbol{g}(\xi) = \text{constant} \end{aligned} \tag{21}$$

Proof: case of $\nu \to 0$

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 15 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.

References

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.

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