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

Omar Chehab, Alexandre Gramfort, Aapo Hyvärinen

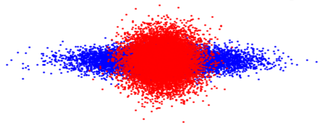
Goal: find the optimal noise distribution for Noise-Contrastive Estimation (NCE)

Approach: Minimize the Asymptotic Variance of NCE w.r.t. the noise distribution.

Contributions:
1. Formulae for the Optimal Noises in popular setups (e.g. noise samples >> data samples)
2. Numerical Validation for Gaussians

Noise-Contrastive Estimation

Noise Contrastive Estimation (NCE) learns a data **model** by contrasting **data** and **noise** samples in a binary classification task



Specifically, a discriminator $D(x)$ is trained with the negative cross-entropy

$$\mathscr{L}_{\mathrm{NCE}}(p_\theta) = -\mathbb{E}_{x\sim p_d}\left[\log D_\theta(x)\right] - \nu\mathbb{E}_{x\sim p_n}\left[\log\left(1-D_\theta(x)\right)\right]$$

and with an infinite amount of samples, recovers the data distribution $p_{\theta^*} = p_d$ where $\theta^* \in \arg\min \mathscr{L}_{\mathrm{NCE}}(p_\theta)$.

Recent work [1, 2] shows the **choice of the noise distribution** is crucial. In practice, it ranges from a default Max-Entropy to adversarially targeting data. But what is the **optimal** noise?

Sample Efficiency

With a limited budget of samples, the parameter is estimated with a statistical error. It is quantified by the MSE [3]

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

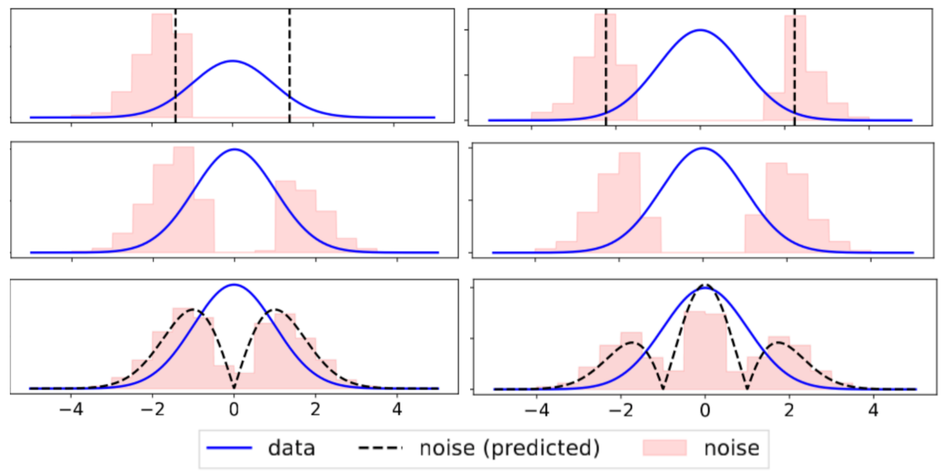
where $\boldsymbol{m}, \boldsymbol{I}$ are a generalized Fisher score mean and covariance. We can choose the noise distribution so that it minimizes that error:

$$p_n^{optimal} \in \arg\min \mathscr{L}_{\mathrm{MSE}}(p_n)$$

What does the Optimal Noise distribution look like?

The task is to estimate the mean or variance of Gaussian data.

Our Theorems	Noise Budget
$p_n \text{ ' } \subseteq \text{ ' } \arg\max\ p_d(x)\ \lVert\nabla_\theta \log p_{\theta^*}\rVert^2$ s.t. $\lVert\nabla_\theta \log p_{\theta^*}\rVert = \text{constant}$	10 %
p_n interpolates both limit cases.	50 %
$p_n \propto p_d(x)\ \lVert\nabla_\theta \log p_{\theta^*}\rVert_{I_F^{-2}}$	90 %



The optimal Noise allocates mass to regions of high data density p_d **and** Fisher score magnitude $\lVert\nabla_\theta \log p_{\theta^*}\rVert$.

It samples data points that are **least robust** under the learnt model, also known as *hard negatives*.

References

[1] Liu et al., Analyzing and Improving the Optimization Landscape of Noise-Contrastive Estimation, *ICLR 2022*.
[2] Rhodes et al., Telescoping Density-Ratio Estimation, *NeurIPS 2020*.
[3] Gutmann et al., Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics, *JMLR 2012*.