REMEDI: Corrective Transformations for Improved Neural Entropy Estimation

Viktor Nilsson [1*†] Anirban Samaddar [2*] Sandeep Madireddy [2] Pierre Nyquist [13]

[1]KTH Royal Institute of Technology [2]Argonne National Laboratory, [3]Chalmers University of Technology and University of Gothenburg

[*]Equal contribution, [†]Corresponding author

Table of Contents

Motivation

- In information theoretic learning, estimation of entropy and other related functionals is fundamental.

Motivation

- In information theoretic learning, estimation of entropy and other related functionals is fundamental.
- Often, the quantity of concern $X \sim \mathbb{P}$ has a continuous distribution (probability measure) in $\mathbb{R}^d$.

Motivation

- In information theoretic learning, estimation of entropy and other related functionals is fundamental.
- Often, the quantity of concern $X \sim \mathbb{P}$ has a continuous distribution (probability measure) in $\mathbb{R}^d$. In other words $\mathbb{P} \ll \lambda$ (the Lebesgue measure) and $\mathbb{P}$ has a density function p_X.

Motivation

- In information theoretic learning, estimation of entropy and other related functionals is fundamental.
- Often, the quantity of concern $X \sim \mathbb{P}$ has a continuous distribution (probability measure) in $\mathbb{R}^d$. In other words $\mathbb{P} \ll \lambda$ (the Lebesgue measure) and $\mathbb{P}$ has a density function p_X.
- For such a quantity we seek to estimate the **differential entropy**

$$H(\mathbb{P}) := \mathbb{E}[-\log p_X(X)].$$

Motivation

- Classical methods for estimating $H(\mathbb{P})$ based on samples $\{x_i\}_{i=1}^n$ from $\mathbb{P}$ include: kernel density estimation (KDE), k-nearest neighbors estimates, methods based on sample spacings.

[1] Wasserman 2004, p. 319.
[2] Pichler et al. 2022.

Motivation

- Classical methods for estimating $H(\mathbb{P})$ based on samples $\{x_i\}_{i=1}^n$ from $\mathbb{P}$ include: kernel density estimation (KDE), k-nearest neighbors estimates, methods based on sample spacings.
- These possess nice asymptotic properties, like consistency, but fail in moderately high dimensions[1].

[1] Wasserman 2004, p. 319.
[2] Pichler et al. 2022.

Motivation

- Classical methods for estimating $H(\mathbb{P})$ based on samples $\{x_i\}_{i=1}^n$ from $\mathbb{P}$ include: kernel density estimation (KDE), k-nearest neighbors estimates, methods based on sample spacings.
- These possess nice asymptotic properties, like consistency, but fail in moderately high dimensions[1].
- KDE may be improved by increasing the model class, e.g. going to Gaussian mixture models (GMMs).

[1]Wasserman 2004, p. 319.
[2]Pichler et al. 2022.

Motivation

- Classical methods for estimating $H(\mathbb{P})$ based on samples $\{x_i\}_{i=1}^n$ from $\mathbb{P}$ include: kernel density estimation (KDE), k-nearest neighbors estimates, methods based on sample spacings.
- These possess nice asymptotic properties, like consistency, but fail in moderately high dimensions[1].
- KDE may be improved by increasing the model class, e.g. going to Gaussian mixture models (GMMs).
- A recent development is entropy estimation with GMMs using gradient-based optimization of a cross-entropy target (KNIFE)[2].

[1] Wasserman 2004, p. 319.
[2] Pichler et al. 2022.

Contribution

This work:

Contribution

This work:

- Shows empirically that the problems of KDE estimates persist in for such GMM estimates.

Contribution

This work:

- Shows empirically that the problems of KDE estimates persist in for such GMM estimates.
- Introduces a deep learning-based correction to the estimates called `REMEDI`, demonstrating good performance in moderate dimensions.

Contribution

This work:

- Shows empirically that the problems of KDE estimates persist in for such GMM estimates.
- Introduces a deep learning-based correction to the estimates called `REMEDI`, demonstrating good performance in moderate dimensions.
- Proves theoretically that it satisfies a desirable consistency property.

Contribution

This work:

- Shows empirically that the problems of KDE estimates persist in for such GMM estimates.
- Introduces a deep learning-based correction to the estimates called `REMEDI`, demonstrating good performance in moderate dimensions.
- Proves theoretically that it satisfies a desirable consistency property.
- Investigates its performance in the Information Bottleneck context.

GMMs fail in moderate dimension

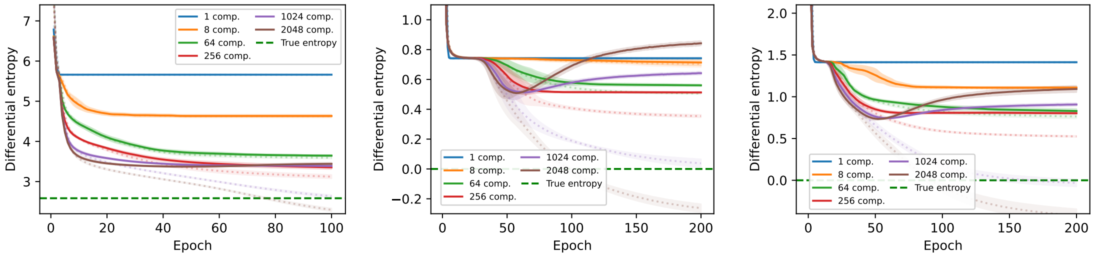


Figure: KNIFE training curves with error bars on 8-dimensional triangle and uniform ball/cube datasets. It is observed that increasing the number of components M for KNIFE leads to overfitting in all datasets.

Method

Task: Estimate $H(\mathbb{P})$.

Method

Task: Estimate $H(\mathbb{P})$.
Idea: Use weakly fitted base models $\mathbb{Q}$ that allow tractable log-likelihoods.

Method

Task: Estimate $H(\mathbb{P})$.
Idea: Use weakly fitted base models $\mathbb{Q}$ that allow tractable log-likelihoods.

- For example normal, Gaussian mixture models, normalizing flows.

Method

Task: Estimate $H(\mathbb{P})$.
Idea: Use weakly fitted base models $\mathbb{Q}$ that allow tractable log-likelihoods.

- For example normal, Gaussian mixture models, normalizing flows.
- The difference between $H(\mathbb{P})$ and the cross-entropy $C(\mathbb{P}||\mathbb{Q})$ is then given by the relative entropy (KL-divergence) $R(\mathbb{P}||\mathbb{Q})$.

Method

Task: Estimate $H(\mathbb{P})$.
Idea: Use weakly fitted base models $\mathbb{Q}$ that allow tractable log-likelihoods.

- For example normal, Gaussian mixture models, normalizing flows.
- The difference between $H(\mathbb{P})$ and the cross-entropy $C(\mathbb{P}||\mathbb{Q})$ is then given by the relative entropy (KL-divergence) $R(\mathbb{P}||\mathbb{Q})$.
- Estimate $R(\mathbb{P}||\mathbb{Q})$ using Donsker-Varadhan's formula.

Method

Donsker-Varadhan's formula: For $\mathbb{P}, \mathbb{Q}$ probability measures on $\mathbb{R}^d$ such that $\mathbb{P} \ll \mathbb{Q}$ we have

$$R(\mathbb{P}||\mathbb{Q}) = \sup_{T \in C_b} \mathbb{E}^{\mathbb{P}}[T] - \log \mathbb{E}^{\mathbb{Q}}[e^T]. \quad (1)$$

Such estimation has been used for mutual information[3].

[3] Belghazi et al. 2018.

Method

Donsker-Varadhan's formula: For $\mathbb{P}, \mathbb{Q}$ probability measures on $\mathbb{R}^d$ such that $\mathbb{P} \ll \mathbb{Q}$ we have

$$R(\mathbb{P}||\mathbb{Q}) = \sup_{T \in C_b} \mathbb{E}^{\mathbb{P}}[T] - \log \mathbb{E}^{\mathbb{Q}}[e^T]. \tag{1}$$

Such estimation has been used for mutual information[3].

- Questions:
  - Can we use empirical estimates in Eq. (1)?

[3] Belghazi et al. 2018.

Method

Donsker-Varadhan's formula: For $\mathbb{P}, \mathbb{Q}$ probability measures on $\mathbb{R}^d$ such that $\mathbb{P} \ll \mathbb{Q}$ we have

$$R(\mathbb{P}||\mathbb{Q}) = \sup_{T \in C_b} \mathbb{E}^{\mathbb{P}}[T] - \log \mathbb{E}^{\mathbb{Q}}[e^T]. \tag{1}$$

Such estimation has been used for mutual information[3].

- Questions:
  - Can we use empirical estimates in Eq. (1)?
  - How to optimize over C_b?

[3] Belghazi et al. 2018.

Method

Donsker-Varadhan's formula: For $\mathbb{P}, \mathbb{Q}$ probability measures on $\mathbb{R}^d$ such that $\mathbb{P} \ll \mathbb{Q}$ we have

$$R(\mathbb{P}||\mathbb{Q}) = \sup_{T \in C_b} \mathbb{E}^{\mathbb{P}}[T] - \log \mathbb{E}^{\mathbb{Q}}[e^T]. \tag{1}$$

Such estimation has been used for mutual information[3].

- Questions:
  - Can we use empirical estimates in Eq. (1)?
  - How to optimize over C_b?
  - Can we obtain theoretical guarantees?

[3] Belghazi et al. 2018.

Loss function

REMEDI loss function: n samples from the data ($\mathbb{P}$) and m independent samples from the base ($\mathbb{Q}$) distribution,

$$\hat{\mathcal{L}}_{\texttt{REMEDI}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} - \log q(x_i)}_{\hat{\mathcal{L}}_{\text{KNIFE}}} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} T(x_i) - \log\left(\frac{1}{m}\sum_{i=1}^{m} e^{T(\tilde{x}_i)}\right)\right)}_{\hat{\mathcal{L}}_{\text{DV}}} \tag{2}$$

Minimizing the loss function (2) implies –

[4] Pichler et al. 2022.

Loss function

REMEDI loss function: n samples from the data ($\mathbb{P}$) and m independent samples from the base ($\mathbb{Q}$) distribution,

$$\hat{\mathcal{L}}_{\text{REMEDI}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} -\log q(x_i)}_{\hat{\mathcal{L}}_{\text{KNIFE}}} \underbrace{-\left(\frac{1}{n}\sum_{i=1}^{n} T(x_i) - \log\left(\frac{1}{m}\sum_{i=1}^{m} e^{T(\tilde{x}_i)}\right)\right)}_{\hat{\mathcal{L}}_{\text{DV}}} \tag{2}$$

Minimizing the loss function (2) implies –

- Minimizing $\hat{\mathcal{L}}_{\text{KNIFE}}$: The cross-entropy between $\mathbb{P}$ and $\mathbb{Q}$

[4] Pichler et al. 2022.

Loss function

$\texttt{REMEDI}$ loss function: n samples from the data ($\mathbb{P}$) and m independent samples from the base ($\mathbb{Q}$) distribution,

$$\hat{\mathcal{L}}_{\texttt{REMEDI}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} -\log q(x_i)}_{\hat{\mathcal{L}}_{\mathrm{KNIFE}}} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} T(x_i) - \log\left(\frac{1}{m}\sum_{i=1}^{m} e^{T(\tilde{x}_i)}\right)\right)}_{\hat{\mathcal{L}}_{\mathrm{DV}}} \tag{2}$$

Minimizing the loss function (2) implies –

- Minimizing $\hat{\mathcal{L}}_{\mathrm{KNIFE}}$: The cross-entropy between $\mathbb{P}$ and $\mathbb{Q}$
- Minimizing $\hat{\mathcal{L}}_{\mathrm{DV}}$: The KL-divergence between $\mathbb{P}$ and $\mathbb{Q}$

[4] Pichler et al. 2022.

Loss function

REMEDI loss function: n samples from the data ($\mathbb{P}$) and m independent samples from the base ($\mathbb{Q}$) distribution,

$$\hat{\mathcal{L}}_{\texttt{REMEDI}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} - \log q(x_i)}_{\hat{\mathcal{L}}_{\text{KNIFE}}} \underbrace{- \left(\frac{1}{n}\sum_{i=1}^{n} T(x_i) - \log\left(\frac{1}{m}\sum_{i=1}^{m} e^{T(\tilde{x}_i)} \right) \right)}_{\hat{\mathcal{L}}_{\text{DV}}} \quad (2)$$

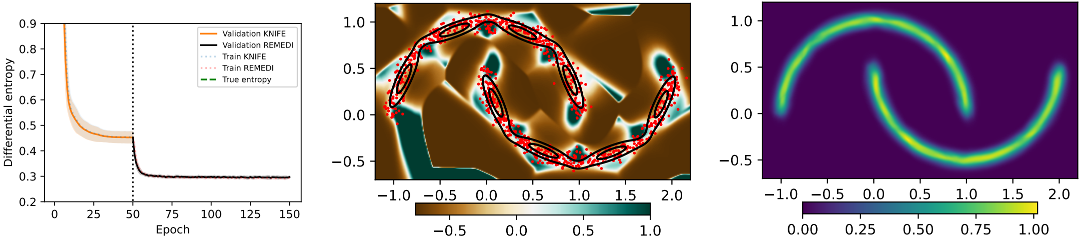
Minimizing the loss function (2) implies –

- Minimizing $\hat{\mathcal{L}}_{\text{KNIFE}}$: The cross-entropy between $\mathbb{P}$ and $\mathbb{Q}$
- Minimizing $\hat{\mathcal{L}}_{\text{DV}}$: The KL-divergence between $\mathbb{P}$ and $\mathbb{Q}$

We select KNIFE[4] as the base distribution.

[4] Pichler et al. 2022.

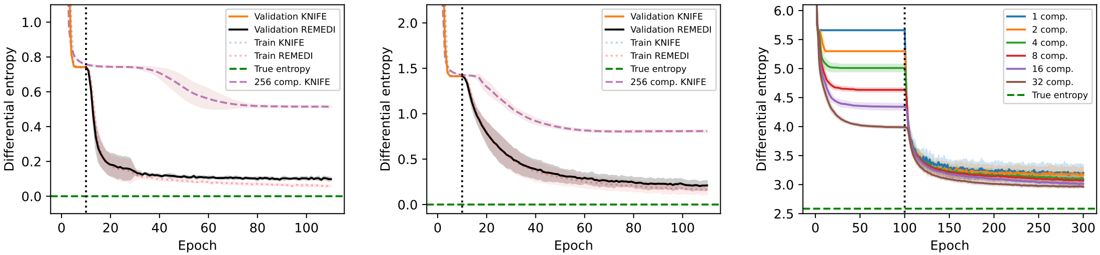
Entropy estimation



(a) Training curve (b) T (image) vs. q (contours). (c) Unnormalized density $qe^T(x)$

Figure: Results on two moons dataset. In the middle we see what direction (positive or negative) REMEDI affects the base distribution. To the right is the unnormalized distribution implied by $q(x)e^{T(x)}$.

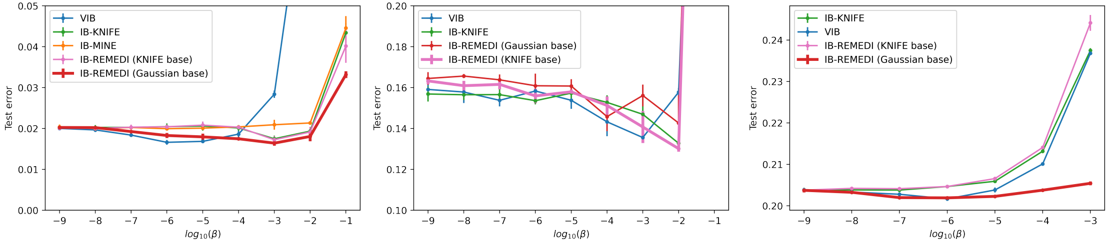
Entropy estimation



(a) (b) (c)

Figure: REMEDI training curves with error bars on 8-dimension uniform ball (a) and cube (b) datasets with 256-comp. KNIFE for reference. (c) The experiment on an 8-dimensional triangle dataset shows the effect of varying the number of components. REMEDI significantly improves the entropy estimation compared to KNIFE.

Information Bottleneck

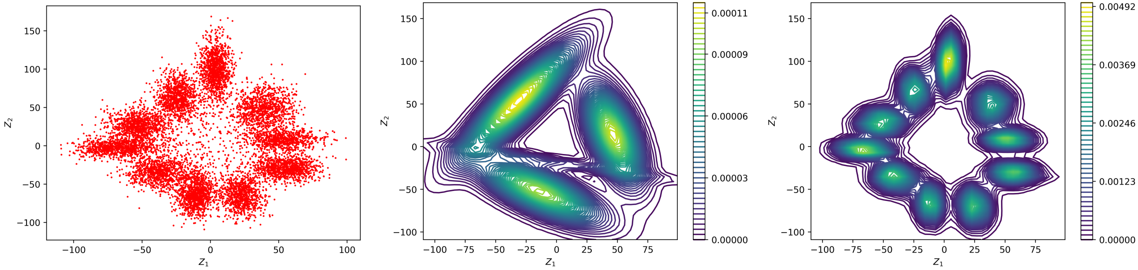


(a) MNIST (b) CIFAR10 (c) ImageNet

Figure: Plot of test error of the Information Bottleneck methods vs β on benchmark image classification datasets with error bars. For most β values, REMEDI outperforms the other methods on MNIST and ImageNet. On CIFAR10, the classification errors are similar for all the methods. However, REMEDI exhibits the lowest classification error across the β values.

Information Bottleneck

Information bottleneck latent space:



(a) Encoder samples (b) KNIFE contours (c) REMEDI contours

Figure: REMEDI marginal distribution of 2-d latent space on MNIST.

Generative capabilities

Rejection sampling: A sample X from $\mathbb{Q}$ is accepted with probability $\phi(X)$, where $\phi(x) = \frac{e^{T(x)}}{\hat{C}}$.

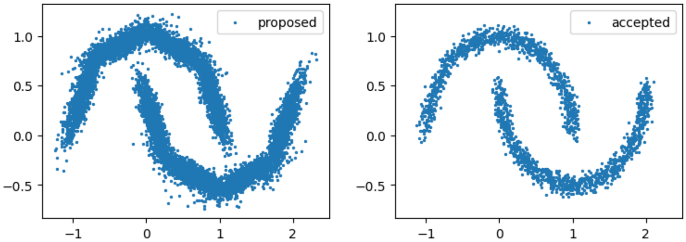


Figure: Left: 10000 proposals from $\mathbb{Q}$. Right: 1989 accepted samples.

Generative capabilities

Langevin diffusion:

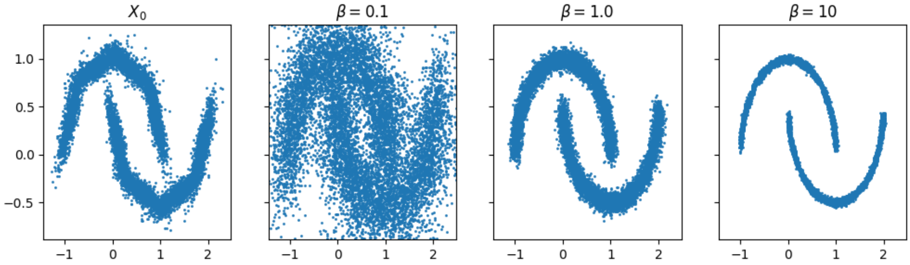


Figure: $\mathbb{Q}$-samples X_0 (leftmost) and X_{t_H} after simulating (3) with different β.

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\beta^{-1}}dW_t, \quad X_0 = x_0, \quad V(x) = -\left(\log q(x) + T(x)\right) \tag{3}$$

Thank you!

References I

- Belghazi, Mohamed Ishmael et al. (2018). “Mutual information neural estimation”. In: *International conference on machine learning*. PMLR, pp. 531–540.
- Pichler, Georg et al. (2022). “A differential entropy estimator for training neural networks”. In: *International Conference on Machine Learning*. PMLR, pp. 17691–17715.
- Wasserman, Larry (2004). *All of statistics: a concise course in statistical inference*. Vol. 26. Springer.