On the Properties of Kullback-Leibler Divergence Between Multivariate Gaussian Distributions

Reporter: Yufeng Zhang
yufengzhang@hnu.edu.cn

Associate professor
College of Computer Science and Electronic Engineering
Hunan University, Changsha, China
2023

Outline

Introduction

statistical divergence

A **statistical divergence** $D : X \times X \to \mathbb{R}^+$ measures the "distance" between probability distributions.

- non-negativity: $D(p, q) \geq 0$
- identity of indiscernibles: $D(p, p) = 0$

statistical distance is stronger, satisfying two extra properties:

- symmetry: $D(p, q) = D(q, p)$
- triangle inequality: $D(p, q) \leq D(p, g) + D(g, q)$

Introduction

Kullback-Leibler divergence

Definition 1

KL divergence *The Kullback-Leibler (KL) divergence between two continuous probability densities* $p(x)$ *and* $q(x)$ *is defined as*

$$KL(p(x)||q(x)) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{1}$$

widely applied in information theory, statistics, and machine learning

Not a proper distance metric

- not symmetric: forward KL divergence $KL(p||q) \to 0$ when reverse $KL(q||p) \to \infty$
- dose not satisfy the triangle inequality

Introduction

Multivariate Gaussian Distribution

Definition 2

Multivariate Gaussian distribution *The probability density function of an n-dimensional Gaussian distribution is given by*

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{2}$$

Here $\boldsymbol{\mu} \in \mathbb{R}^n$ is the mean and $\boldsymbol{\Sigma} \in \boldsymbol{S}^n_{++}$ is the covariance matrix, where $\boldsymbol{S}^n_{++}$ is the space of symmetric positive definite $n \times n$ matrices.

one of the most important distributions

- widely used in many fields

Introduction

KL divergence between multivariate Gaussian distributions

Definition 3

*The KL divergence between two n-dimensional Gaussians $KL(\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))$ has the following closed form (***Pardo 2018***)*

$$\frac{1}{2}\left\{\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \mathrm{Tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - n\right\} \tag{3}$$

where the logarithm is taken to base e and Tr is the trace of matrix.

not symmetric and does not satisfy the triangle inequality either.

$$\text{forward } KL(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(0, I)) = \frac{1}{2}\left\{-\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\boldsymbol{\mu} - n\right\} \tag{4}$$

$$\text{reverse } KL(\mathcal{N}(0, I)||\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2}\left\{\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - n\right\} \tag{5}$$

Introduction

Theoretical Research Questions

For any n-dimensional multivariate Gaussian distributions $\mathcal{N}_1, \mathcal{N}_2$ and $\mathcal{N}_3$

1 **The approximate symmetry of small KL divergence between Gaussians:** when $KL(\mathcal{N}_1||\mathcal{N}_2) \leq \varepsilon$, $KL(\mathcal{N}_2||\mathcal{N}_1) \leq ?$

2 When $KL(\mathcal{N}_1||\mathcal{N}_2) \geq M$, $KL(\mathcal{N}_2||\mathcal{N}_1) \geq ?$

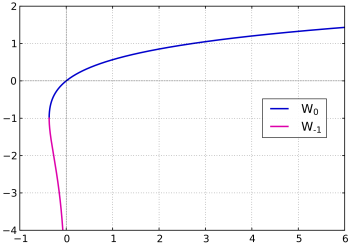
3 **Relaxed triangle inequality:** when $KL(\mathcal{N}_1||\mathcal{N}_2) \leq \varepsilon_1$, and $KL(\mathcal{N}_2||\mathcal{N}_3) \leq \varepsilon_2$, $KL(\mathcal{N}_1||\mathcal{N}_3) < ?$

Definition

Definition 4

Lambert W Function (Lambert 1758; Corless et al. 1996). *The reverse function of function $y = xe^x$ is called Lambert W function $y = W(x)$.*

When $x \in \mathbb{R}$, W is a multivalued function with two branches W_0, W_{-1}, where W_0 is the principal branch (also called branch 0) and W_{-1} is the branch -1.



Main Results

Approximate symmetry of KL divergence between Gaussians

Theorem 1

Approximate symmetry of KL divergence between Gaussians *For any two n-dimensional Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, if $KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \leq \varepsilon (\varepsilon \geq 0)$, then*

$$
\begin{aligned}
& KL(\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) \\
\leq & \frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - \log\frac{1}{-W_0(-e^{-(1+2\varepsilon)})} - 1\right) \\
= & \varepsilon + 2\varepsilon^{1.5} + O(\varepsilon^2) \text{ (for small } \varepsilon)
\end{aligned}
$$

The supremum is attained when the following two conditions hold.

(1) There exists only one eigenvalue λ_j of $B_2^{-1}\boldsymbol{\Sigma}_1(B_2^{-1})^\top$ or $B_1^{-1}\boldsymbol{\Sigma}_2(B_1^{-1})^\top$ equal to $-W_0(-e^{-(1+2\varepsilon)})$ and all other eigenvalues λ_i $(i \neq j)$ are equal to 1, where $B_1 = P_1 D_1^{1/2}$, P_1 is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}_1$, $D_1 = diag(\lambda_1, \ldots, \lambda_n)$ whose diagonal elements are the corresponding eigenvalues, B_2 is defined in the similar way as B_1 except on $\boldsymbol{\Sigma}_2$.

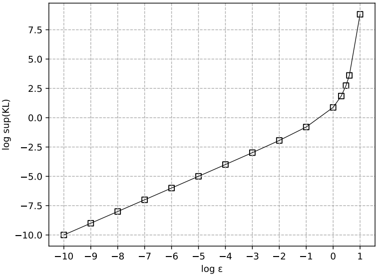
(2) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Approximate symmetry of KL divergence between Gaussians

Remarks

The supremum in Theorem 1 has the following properties.

1. The supremum is small (resp. 0) when ε is small (resp. 0)
2. The supremum increases rapidly when $\varepsilon > 2$
3. It needs strict conditions to reach the supremum
4. The bound is independent of the dimension n. This is a critical property in high-dimensional problems.



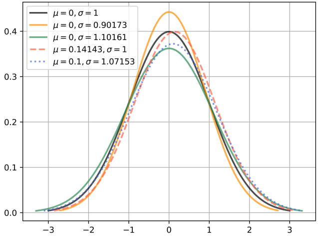
Approximate symmetry of KL divergence between Gaussians

Toy Examples

$\mathcal{N}_0(0,1)$: standard Gaussian distribution (in black). $\mathcal{N}_i$ $(1 \leq i \leq 4)$: $KL(\mathcal{N}_i||\mathcal{N}_0) = 0.01$.
$\mathcal{N}_1$ has the maximized reverse KL divergence:

$$KL(\mathcal{N}_0||\mathcal{N}_1(0, 0.90173^2)) \approx 0.01148 \approx \frac{1}{2}\left(\frac{1}{-W_0(-e^{-(1+2\times 0.01)})} - \log\frac{1}{-W_0(-e^{-(1+2\times 0.01)})} - 1\right)$$

$KL(\mathcal{N}_0||\mathcal{N}_2(0, 1.10161^2)) \approx 0.00879$
$KL(\mathcal{N}_0||\mathcal{N}_3(0.14143, 1)) \approx 0.01$
$KL(\mathcal{N}_0||\mathcal{N}_4(0.1, 1.07153^2)) \approx 0.00892.$



Main Results

Theorem 2

For any two n-dimensional Gaussians $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, if $KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \geq M$, then

$$\begin{aligned} & KL(\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) \\ \geq & \frac{1}{2}\left\{\frac{1}{-W_{-1}(-e^{-(1+2M)})} - \log\frac{1}{-W_{-1}(-e^{-(1+2M)})} - 1\right\} \end{aligned}$$

Theorem 1 and Theorem 2 form a duality

- can be proved in the similar way
- deduce each other

Main Results

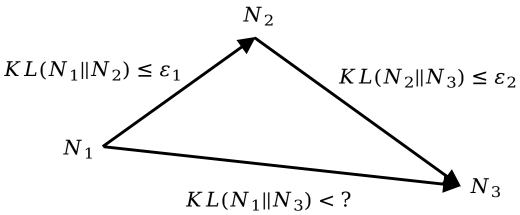
Relaxed triangle inequality

Theorem 3

Relaxed triangle inequality *For any three n-dimensional Gaussians* $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, *and* $\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$, *if* $KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \leq \varepsilon_1, KL(\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)||\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)) \leq \varepsilon_2$, *then*

$$KL(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)) < 3\varepsilon_1 + 3\varepsilon_2 + 2\sqrt{\varepsilon_1\varepsilon_2} + o(\varepsilon_1) + o(\varepsilon_2)$$

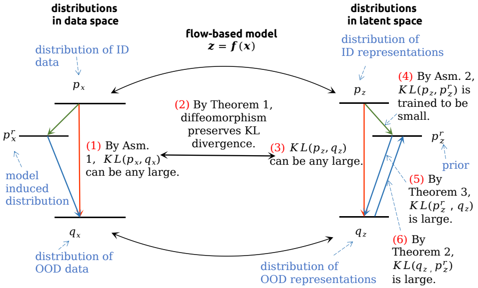
The bound is also dimension-free.



Applications

1: Deep anomaly detection (motivating application)

Research questions:
How to detect anomaly data using flow-based models?
Theorem 1, 2, and 3 provide solid theoretical basis for deep anomaly detection method. See[1] for details.



[1] Yufeng Zhang et al. (2023). "Kullback-Leibler Divergence-Based Out-of-Distribution Detection with Flow-Based Generative Models". In: *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14. DOI: 10.1109/TKDE.2023.3309853.

Applications

Approximate Symmetry of Small KL divergence

How Theorem 1 can help us?

1 Minimizing one of forward and reverse KL divergences also bounds another.

2 We can exchange forward and reverse KL divergences for small ε.

Applications:

1 **Extending theoretical guarantee for discrete policies to continuous Gaussian policy in offline reinforcement learning**. See (Nair et al. 2021).

2 **Bridging research on sample complexity of learning Gaussian distributions**. Current work derive sample complexity using forward and reverse KL divergence separately. We can eliminate such difference. See (Ashtiani et al. 2020; Bhattacharyya et al. 2022).

3 **Bringing new insights to existing reinforcement learning algorithm**. We can exchange forward and reverse KL divergence in MPO algorithm. See (Abdolmaleki et al. 2018).

Applications

Relaxed Triangle Inequality

The relaxed triangle inequality (Theorem 3) can extend one-step robustness guarantee to multiple steps for safe reinforcement learning (Liu et al. 2022).

Related Work

- No existing work focus on the similar properties of KL divergence between Gaussians
- estimation of divergences
  Wang, Kulkarni, and Verdu 2009; Nguyen, Wainwright, and Jordan 2010; Moon and Hero 2014; Rubenstein et al. 2019
- other divergences in different contexts
  Gulrajani et al. 2017; Donnat, Marti, and Very 2016; Abou-Moustafa and Ferrie 2012, Pardo 2018.
- different from existing generalized Pythagoras inequalities

Conclusion

For any n-dimensional multivariate Gaussian distributions $\mathcal{N}_1, \mathcal{N}_2$ and $\mathcal{N}_3$

1. **The approximate symmetry of small KL divergence between Gaussians**
2. **Relaxed triangle inequality**
3. Applications in deep anomaly detection, reinforcement learning, and sample complexity research.

References

Abdolmaleki, Abbas et al. (2018). "Maximum a Posteriori Policy Optimisation". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=S1ANxQW0b.

Abou-Moustafa, Karim T and Frank P Ferrie (2012). "A note on metric properties for some divergence measures: The Gaussian case". In: *Asian Conference on Machine Learning*. PMLR, pp. 1–15.

Ashtiani, Hassan et al. (2020). "Near-Optimal Sample Complexity Bounds for Robust Learning of Gaussian Mixtures via Compression Schemes". In: *J. ACM* 67.6. ISSN: 0004-5411. DOI: 10.1145/3417994. URL: https://doi.org/10.1145/3417994.

Bhattacharyya, Arnab et al. (2022). "Learning Sparse Fixed-Structure Gaussian Bayesian Networks". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 9400–9429. URL: https://proceedings.mlr.press/v151/bhattacharyya22b.html.

Corless, Robert M et al. (1996). "On the Lambert W function". In: *Advances in Computational mathematics* 5.1, pp. 329–359.

Cover, Thomas M and Joy A Thomas (2012). *Elements of information theory*. John Wiley & Sons.

Donnat, Philippe, Gautier Marti, and Philippe Very (Jan. 2016). "Toward a Generic Representation of Random Variables for Machine Learning". In: *Pattern Recogn. Lett.* 70.C, 24–31. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2015.11.004. URL: https://doi.org/10.1016/j.patrec.2015.11.004.

Erven, Tim van and Peter Harremos (2014). "Rényi Divergence and Kullback-Leibler Divergence". In: *IEEE Transactions on Information Theory* 60.7, pp. 3797–3820. DOI: 10.1109/TIT.2014.2320500.

Thanks