SELF-SUPERVISED LEARNING ENCODES UNCERTAINTY

Miguel De Llanza Varona[1], **Ryan Singh**[1], **Christopher L. Buckley**[1,2]
[1]School of Engineering and Informatics, University of Sussex, UK
[2]VERSES AI Research Lab, Los Angeles, California, 90016, USA
{md623,rs773}@sussex.ac.uk

ABSTRACT

Practitioners have become aware that self-supervised learning techniques using multiple views (created through augmentation) outperform reconstruction-based methods on downstream tasks. Intuitive arguments suggest this is due to the dimensionality of the observation space. Another theoretical line of attack is through work on provable disentanglement under the assumption original image is recoverable from each view. We extend these arguments to the case where the assumptions are dropped. To do this we connect to traditional statistical theory by casting SSL as a method for learning sufficient statistics. This allows us to show when exact recoverability is not possible SSL representations are (information theoretically) equivalent to posterior distributions. We demonstrate, in a toy model with known data generating process, even as the original data becomes corrupted by noise the SSL representations remain correlated with the posterior distribution. We further demonstrate that the representations specifically correlate with the posterior variance, indicating uncertainty is being encoded. We believe this viewpoint can shed new light on the question on when reconstruction-methods fail, for example, when likelihoods are difficult to represent but sampling is cheap and sufficient statistics are simple.

1 INTRODUCTION

One of the main goals of representation learning is to encode the latent structure causing the data-generation process. In recent years, self-supervised learning (SSL) has emerged as the main approach to learning useful representations in unsupervised settings (Assran et al., 2023; Bardes et al., 2022; Ermolov et al., 2021; Liu et al., 2022; Zbontar et al., 2021; Chen et al., 2020). SSL representations can generalize well across multiple downstream tasks, even reaching the supervised baseline in some cases.

As opposed to reconstruction-based methods, SSL doesn't encode an explicit generative model; instead, it leverages some underlying structure in the data to create a supervisory signal that highlights certain features over others. One popular strategy is to generate multiple views from the same observation to learn a joint representation that captures the redundant information while discarding view-specific information (Federici et al., 2020; Hjelm et al., 2019; Oord et al., 2019). In this process, multi-view SSL exploits the invariant information across views to make some part of the input predictable from another part.

The failure of reconstruction-based models is often attributed to the difficulty of building "good-enough" generative models of high-dimensional data. Balestriero & LeCun (2024); Shwartz-Ziv et al. argue that a main challenge in generative modelling is to capture the relevant directions of variance in the data, or "perceptual features", which often misalign with those in service of reconstruction. In the context of variational inference, it has been shown that maximizing the ELBO is not a guarantee of learning good posterior approximations (Zhao et al. (2018; 2017)). This posits some doubts about the role of reconstruction in learning useful representations. Similarly, it has been suggested that reconstruction-based objectives are ill-posed when the data is generated from a lower-dimensional manifold Loaiza-Ganem et al. (2024). More generally, constraints such as

limited encoding capacity or misspecified likelihood models can hinder the learning of useful representations in latent variable models.

In this work, we rely on traditional statistical theory to cast SSL as a method for learning sufficient statistics. We formalize the concept of minimal sufficient statistics in the context of multiple views and their relation to traditional sufficiency. We show theoretically how in unconstrained scenarios multi-view SSL methods recover the full posterior distribution up to a deterministic transformation. Finally, we design two toy models to illustrate these theoretical claims.

1.1 Related work

SSL and inversion of the data-generation process. We build up on previous work on contrastive learning and implicit data modelling. They show that under certain assumptions, SSL recovers the data-generation process (Zimmermann et al., 2022). Similarly, (Kügelgen et al., 2022) demonstrate that SSL can disentangle irrelevant from relevant information by implicitly inverting the data-generation process. This line of work shows that SSL learn an implicit generative model of the data.

Likelihood-free inference and contrastive learning. Likelihood-free inference rely on a simulator to approximate the posterior distribution when the likelihood is intractable (Thomas et al. (2020); Gutmann et al. (2018); Hermans et al. (2020); Greenberg et al. (2019)). Zimmermann et al. (2022) unify different likelihood-free inference methods under a contrastive learning framework.

Bayesian Inference and SSL. Recent work has examined the relation between Bayesian inference and contrastive learning. Walker et al. (2023); Aitchison & Ganev (2023) rely on a recognition parametrized to learn a latent variable model without explicitly defining a generative model.

Multi-view SSL. Federici et al. (2020) explore multi-view SSL from an information theory perspective. In particular, they extend the information bottleneck hypothesis to the multi-view setting and define minimal sufficient statistics as superfluous and predictive information respectively.

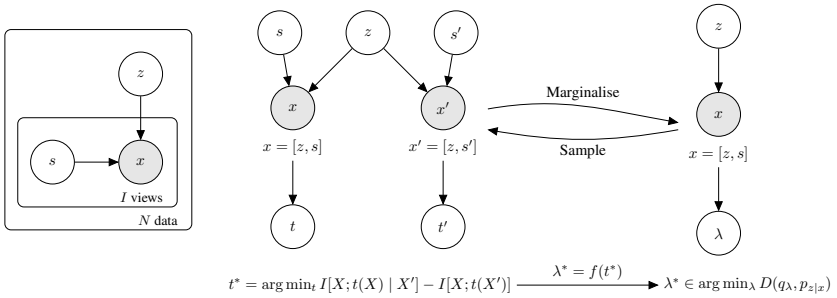


Figure 1: Schematic representation of the correspondence between multiview self-supervised learning and Bayesian Inference. The multiview representation can be converted to the local latent variable model by marginalising over augmentations, while in the other direction we can produce multiple views by sampling with the latent variable fixed. Our main result uses the notion of sufficiency to show there exists a deterministic mapping between the representation t and the posterior parameters λ.

2 Background

2.1 Sufficiency

Classically, sufficient statistics are functions of the data which maximally preserve the ability to discriminate between a set of hypotheses, while minimal sufficient statistics aim to discard any information not useful for this task. Formally, for a set of models $\{p_\theta(x) \mid \theta \in \Theta\}$ a function $t(x)$ is sufficient if $p(x \mid t, \theta) = p(x \mid t)$ for every $\theta \in \Theta$. While t is minimal if for any other statistic s there exists a function $f(s) = t$. This definition immediately implies that minimal sufficient statistics satisfy a type of uniqueness or universal property which we will apply to show there must be a mapping between SSL representations and posterior parameters.

There is a tight link between information theory and sufficiency. Indeed Kullback & Leibler (1951), building on Halmos & Savage (1949), introduced their eponymous divergence to give an alternate characterisation of sufficiency. They showed, in modern notation, that t is sufficient iff $I[X;\theta] = I[T;\theta]$. Which follows from an application of the data processing inequality. Similar arguments show that if t is sufficient and $I[X;T] = I[X;\theta]$ then T is minimal.

These qualities are useful for analysing modern self-supervised techniques. For example, it has previously pointed out by Shamir et al. (2010) that the Information Bottleneck method:

$$\mathcal{L}_{IB}[t] = I[X;T] - \beta I[T;\theta] \tag{1}$$

generalises classical sufficient statistics. Intuitively, the term $I[T;\theta]$ chooses functions that capture information about the parameter of interest, while $I[X;T]$ favours compression, or minimality.

2.2 SSL Objectives

Self-supervised methods that use augmentations are often related to either the information bottleneck or to the infomax principle. Most prominently the Multiview-IB (Federici et al., 2020):

$$\mathcal{L}_{mIB}[t] = I[X;T \mid X'] - \beta I[T;X'] \tag{2}$$

which aim to capture the notion of sufficiency through views X, X' of the data. Federici et al. (2020) showed under the assumption each view is 'redundant' with respect to the other and T is sufficient for (X, X') (i.e. $I[X;X' \mid T] = 0$) then T is sufficient for θ.

More generally infomax methods simply aim to maximise $I[T;\theta]$, in this paper we will mainly consider the InfoNCE objective (Oord et al., 2019; Chen et al., 2020): z for some kernel function k, where pos samples are drawn from a common latent and neg samples are drawn independently from the joint, which is a lower bound on the the mutual information $I[T;\theta] \geq -\mathcal{L}_{NCE}$. Finally Shwartz-Ziv et al. also studied how VICReg (Bardes et al., 2022) can be seen as approximate infomax under certain assumptions.

2.3 Bayes Optimality

On the other hand optimality properties of the bayesian posterior in information processing (Zellner, 1988), optimal control (Striebel, 1965), point estimation (Lehmann & Casella, 1998) and decision theory (Bernardo & Smith, 2009) often rely on the notion the posterior distribution contains the 'optimal' amount of information about the data with respect to the prior. Here we show a similar result specialised to the setting of finite dimensional variational inference. We show Lemma 1 that if $Q = \{q_\lambda(\theta)\}_{\lambda \in \Lambda}$ is a variational family, and $t : X \to \Lambda$ is the amortised posterior mapping, then t is a minimal sufficient statistic.

To connect these notions, similarly to Kügelgen et al. (2022); Zimmermann et al. (2022), we introduce a correspondence between latent variable models and self-supervised learning with augmentations Figure 1. Specifically, we identify the underlying distribution before augmentations with draws from a prior distribution $\theta_{1:n} \sim \pi$, while the views are samples from a likelihood $x_i, x_i' \sim l(x \mid \theta_i)$.

This likelihood function can be explicitly related to an augmentation process. We assume there is a conditional density $s \sim f(x \mid \theta, s)$ indexed by an augmentation parameter s, where s is also sampled from some distribution P_s. The likelihood function can then be identified with the marginalised distribution $l(x \mid \theta) = \mathbb{E}_{p_s}[f(x \mid \theta, s)]$.

3 Theory

Our main result is to show in certain situations multi-view SSL methods recover the posterior distribution up to a deterministic transformation. We need the following assumptions:

Assumption 1. *For any $\theta_1, \theta_2 \in \Theta$ there exists $x, x' \in X$ such that the likelihood ratio's are not equal:*

$$\frac{l(x' \mid \theta_1)}{l(x' \mid \theta_2)} \neq \frac{l(x \mid \theta_1)}{l(x \mid \theta_2)}$$

Assumption 2. *There is a statistical manifold $Q = \{q_\lambda(\theta) \mid \lambda \in \Lambda\}$, with $\Lambda \subset \mathbb{R}^d$ and a unique continuous amortisation map $x \mapsto \lambda$ such that $p(\theta \mid x) = q_\lambda(\theta)$*

Proposition 1 (Multiview Posteriors)**.** *Given the assumptions, suppose t_β minimises Equation* (2) *for some prior and augmentation procedure (π, f) there exists β and an invertible function $g_\pi : T \to \Lambda$ such that $p(\theta \mid x) = q_{g_\pi(t_\beta(x))}(\theta)$.*

Proof Sketch First, Assumption 2 guarantees that the amortisation map is a minimal sufficient statistic. So if we can show t_β is also a minimal sufficient statistic then we can gaurantee the existence of f_π. Minimal sufficiency of t_β *for the other view*, X', follows from the fact the Multiview Bottleneck Equation (2) is a generalisation of the information bottleneck. Assumption 1 guarantees that any information useful for predicting X' is also useful for discriminating θ hence t_β is minimal sufficient for θ.

We extend this result to InfoNCE by exploiting the equality condition of Donsker-Varadhan representation which shows that the optimal t is again (X, X') sufficient. This allows us to use the majority of the previous argument, however since we do not have minimality, the function is not invertible. This reflects a limitation of infomax methods, which do not penalise superfluous information, as discussed by Federici et al. (2020). In the case of exact recoverability (Kügelgen et al., 2022), this reflects the assumption that the dimension of the latent variable is known a-priori.

Proposition 2 (InfoNCE Posteriors)**.** *Given the assumptions, suppose t minimises Equation* (12) *for some prior and augmentation procedure (π, f), there exists a function $g_\pi : T \to \Lambda$ such that $p(\theta \mid x) = q_{g_\pi(t_\beta(x))}(\theta)$.*

4 EXPERIMENTS

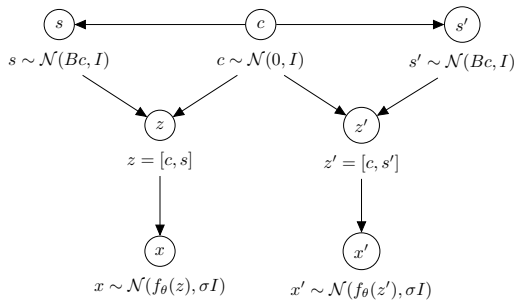
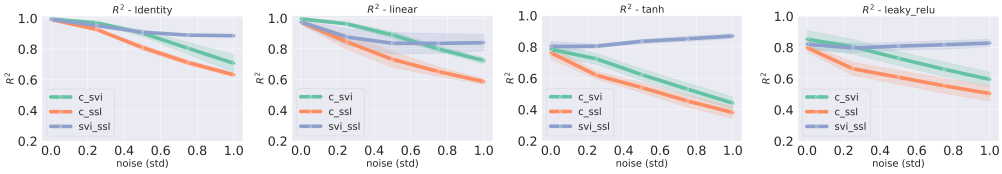


Figure 2: Generative process toy models.

We based our toy models on Kügelgen et al. (2022) to show that InfoNCE approximates, up to a linear transformation, the sufficient statistics learned by stochastic variational inference (SVI) with access to the ground truth generative model. We choose a multivariate normal Gaussian with an identity covariance matrix as our mean-field variational posterior. Figure 2 describes the multiview generative process. We assume that the parameter of interest is c, and both s and s' contain the irrelevant (or view-specific) information to be discarded by InfoNCE. Following Figure 1, SVI marginalizes over s and s' to get rid of the irrelevant information in the observations, similarly as InfoNCE does. Thus, the aim is to infer c from noisy observations x and x'. Throughout both experiments, we set the dimensionality of c, s, and s' to 5 and we define f_θ as an MLP (see B for details).

4.1 TOY MODEL 1

In the first experiment, we set the latent dimension of our InfoNCE model to the same size as μ in SVI (5). For different noise scales σ, we measure the R^2, up to linear transformation, between: i) SVI_μ and c; ii) InfoNCE and c; and iii) SVI_μ and InfoNCE. By gradually increasing the noise, we expect i) and ii) to degrade accordingly, while iii) remains constant. This would indicate that both InfoNCE and SVI converge to the same parameter estimates, thus encoding the same information.



(a) (b) (c) (d)

Figure 3: Linear R^2 between the true relevant parameter c and InfoNCE and SVI_μ (orange and green lines respectively). The blue line shows the linear R^2 between InfoNCE and SVI_μ. We test this for different f_θ: a) the identity; b) a linear transformation; c) linear transformation plus a tanh function; d) linear transformation plus a leaky relu function. The same pattern emerges across all configurations: the predictability between SVI_μ and InfoNCE remains stable as the noise scales up.

Indeed, Figure 3 shows that the predictability between InfoNCE and SVI_μ is noise-robust as their predictability remains fairly stable. In addition, their predictability of c degrades at a similar rate. This result suggests that both methods capture similar information about c.

4.2 Toy Model 2

In the next experiment, we test the ability of SSL to encode uncertainty. This is key to evaluate whether SSL learns to model uncertainty as part of the process of recovering sufficient statistics. For this purpose, we introduce uncertainty in the generative process by partially collapsing information in the observation space through a bottleneck in f_θ. In particular, we first sample an observation $x \sim \mathcal{N}(z, I\sigma)$ which is then non-linearly transformed into a 5-dimensional space (instead of 10): $dim(f_\theta) = 5$. We set $\sigma = 0.001$ to reduce the noise interference and ease the task. Since this data-generation process creates some intrinsic uncertainty about the true parameters, we measure the R^2, up to linear transformation, between the whole set of parameters learned by $SVI_{\mu\sigma}$ (means and variances) and the latent variables of InfoNCE. In particular, we sweep across different latent sizes to see whether extra latent dimensions encode any uncertainty modelled by SVI.

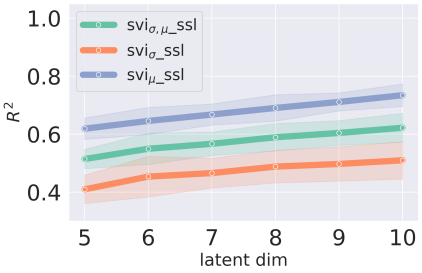


Figure 4: R^2 score between SVI and SSL as the number of SSL latent dimensions increases. We include the σ parameters of the SVI to test whether extra SSL latent variables encode any uncertainty.

Figure 4 shows that as the number of latent dimensions increases, SSL predicts more accurately the full posterior distribution learned by SVI (green line). This result suggests that SSL can encode a richer representation by modelling uncertainty. As there are no disentanglement guarantees in the SSL latent space, it is expected to see an increase in the mean estimates as well (i.e., a latent dimension might encode mean and variance information), shown by the blue line. We speculate that the increased complexity introduced by the bottleneck in the MLP is the reason why, for a latent dimension of 5, the difference in R^2 between SVI and InfoNCE is significant (blue lines in Figure 3d and Figure 4 respectively).

5 Limitations and Discussion

While we aimed to elucidate the connection between SSL, sufficient statistics and posteriors, we required several restrictive assumptions. Importantly, we assumed that the task was simple enough i.e. finite dimensional posteriors exist, and that the encoders were expressive enough to recover sufficient statistics. We realise the true advantage of SSL based methods might be in the case where neither of these assumptions hold. On the other hand, a useful feature of Bayesian inference is it's sensitivity to different priors. This could be particularly important in control tasks where epistemic

uncertainty is constantly updated in the face of new data. Finally, our experiments were limited to toy models where we could extract ground truth posterior estimates, it is an unclear, but potentially interesting question, how one would extract uncertainty estimates without access to the data generating process.

ACKNOWLEDGMENTS

This work was supported by The Leverhulme Trust through the be.AI Doctoral Scholarship Programme in biomimetic embodied AI and VERSES AI.

REFERENCES

Laurence Aitchison and Stoil Ganev. InfoNCE is variational inference in a recognition parameterised model, August 2023. URL http://arxiv.org/abs/2107.02495. arXiv:2107.02495 [stat].

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture, April 2023. URL http://arxiv.org/abs/2301.08243. arXiv:2301.08243 [cs].

Randall Balestriero and Yann LeCun. Learning by Reconstruction Produces Uninformative Features For Perception, February 2024. URL http://arxiv.org/abs/2402.11337. arXiv:2402.11337.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, January 2022. URL http://arxiv.org/abs/2105.04906. arXiv:2105.04906.

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, September 2009. ISBN 978-0-470-31771-6. Google-Books-ID: 11nSgIcd7xQC.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, November 2020. URL https://proceedings.mlr.press/v119/chen20j.html. ISSN: 2640-3498.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for Self-Supervised Representation Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3015–3024. PMLR, July 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html. ISSN: 2640-3498.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning Robust Representations via Multi-View Information Bottleneck, February 2020. URL http://arxiv.org/abs/2002.07017. arXiv:2002.07017.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic Posterior Transformation for Likelihood-Free Inference. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2404–2414. PMLR, May 2019. URL https://proceedings.mlr.press/v97/greenberg19a.html. ISSN: 2640-3498.

Michael U. Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, March 2018. ISSN 1573-1375. doi: 10.1007/s11222-017-9738-6. URL https://doi.org/10.1007/s11222-017-9738-6. TLDR: This work finds that classification accuracy can be used to assess the discrepancy between simulated and observed data and the complete arsenal of classification methods becomes thereby available for inference of intractable generative models.

Paul R. Halmos and L. J. Savage. Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics. *The Annals of Mathematical Statistics*, 20(2):225–241, June 1949. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177730032. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-20/issue-2/

Application-of-the-Radon-Nikodym-Theorem-to-the-Theory-of/10.1214/aoms/1177730032.full. Publisher: Institute of Mathematical Statistics.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with Amortized Approximate Ratio Estimators. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4239–4248. PMLR, November 2020. URL https://proceedings.mlr.press/v119/hermans20a.html. ISSN: 2640-3498.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, February 2019. URL http://arxiv.org/abs/1808.06670. arXiv:1808.06670.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. ISSN 0003-4851. URL https://www.jstor.org/stable/2236703. Publisher: Institute of Mathematical Statistics.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style, January 2022. URL http://arxiv.org/abs/2106.04619. arXiv:2106.04619.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, NY Berlin Heidelberg, 2nd ed. 1998 edition edition, August 1998. ISBN 978-0-387-98502-2.

Xin Liu, Zhongdao Wang, Ya-Li Li, and Shengjin Wang. Self-Supervised Learning via Maximum Entropy Coding. *Advances in Neural Information Processing Systems*, 35:34091–34105, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dc709714c52b35f2f34aca2a92b06bc8-Abstract-Conference.html.

Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections, April 2024. URL https://arxiv.org/abs/2404.02954v2.

Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, January 1997. doi: 10.1098/rsta.1933.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1933.0009. Publisher: Royal Society.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748.

Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696–2711, June 2010. ISSN 0304-3975. doi: 10.1016/j.tcs.2010.04.006. URL https://www.sciencedirect.com/science/article/pii/S030439751000201X.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G J Rudner, and Yann LeCun. An Information-Theoretic Perspective on Variance-Invariance-Covariance Regularization.

Charlotte Striebel. Sufficient statistics in the optimum control of stochastic systems. *Journal of Mathematical Analysis and Applications*, 12(3):576–592, December 1965. ISSN 0022-247X. doi: 10.1016/0022-247X(65)90027-2. URL https://www.sciencedirect.com/science/article/pii/0022247X65900272.

Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U. Gutmann. Likelihood-free inference by ratio estimation, September 2020. URL http://arxiv.org/abs/1611.10242. arXiv:1611.10242.

William I. Walker, Hugo Soulat, Changmin Yu, and Maneesh Sahani. Unsupervised representation learning with recognition-parametrised probabilistic models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 4209–4230. PMLR, April 2023. URL https://proceedings.mlr.press/v206/walker23a.html. ISSN: 2640-3498.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12310–12320. PMLR, July 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html. ISSN: 2640-3498.

Arnold Zellner. Optimal Information Processing and Bayes's Theorem. *The American Statistician*, 42(4):278–280, 1988. ISSN 0003-1305. doi: 10.2307/2685143. URL https://www.jstor.org/stable/2685143. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards Deeper Understanding of Variational Autoencoding Models, February 2017. URL http://arxiv.org/abs/1702.08658. arXiv:1702.08658 [cs].

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information Maximizing Variational Autoencoders, May 2018. URL http://arxiv.org/abs/1706.02262. arXiv:1706.02262.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive Learning Inverts the Data Generating Process, April 2022. URL http://arxiv.org/abs/2102.08850. arXiv:2102.08850 [cs].

A Appendix

You may include other additional sections here.

A.1 Bayesian Inference as Sufficient Statistics

General setup and notations, we will work with a statistical manifold $P = \{p_\theta(x) \mid \theta \in \Theta\}$ of densities with respect to a common measure space (X, Σ_X, μ), and Θ is assumed to be a lebesgue measurable subset of euclidean space. Often we will also have a π which we assume to be absolutely continuous with respect to the lebesgue measure. We will be considering functions $t : X \to Y$ which are $(\mathcal{X}, \Sigma_x)$ measurable. Finally as in variational inference we will consider a second statistical manifold $Q = \{q_\lambda(\theta) \mid \lambda \in \Lambda\}$ we again assume Λ is lebesgue measurable.

For our purposes we will start from the conclusion of the neyman pearson factorisation theorem (Neyman et al., 1997).

Definition A.1. *Let $P = \{p_\theta(x) \mid \theta \in \Theta\}$ be a set of density functions with respect to a common measure μ on X. Then a μ measurable function $t : X \to Y$ is sufficient for P if $p_\theta(x) = f_\theta(t(x))g(x)$ for all $p_\theta \in P$*

Definition A.2. *A sufficient statistic is minimal if for any other sufficient statistic, $s : X \to Z$ there exists $f : Z \to Y$, such that $t = f \circ s$, i.e. $T = f(S)$*

As a straightforward consequence the posterior density $p_\pi(\theta \mid x) = p_\pi(\theta \mid t(x))$, which we assume exists.

Lemma 1. *Suppose $Q = \{q_\lambda(\theta) : \lambda \in \Lambda\}$ is a statistical manifold indexed by Λ and given a prior π define the map*

$$\begin{aligned} t :& X \to \Lambda \\ & x \mapsto \lambda : p_\pi(\theta \mid x) = q_\lambda(\theta) \quad \pi\ a.s. \end{aligned}$$

if it exists and is continuous (e.g. a neural network), then t is a sufficient statistic, further if for any x there is a unique λ then t is minimal.

Proof. Measurability follows from continuity. Next for sufficency, note $p_\pi(x, \theta) = p_\pi(\theta \mid x)p(x) = q_\lambda(\theta)p(x)$ where $p_\pi(x) = \int p_\pi(x \mid \theta)d\pi$ therefore $p(x \mid \theta) = \frac{1}{\pi(\theta)} q_\lambda(\theta)p_\pi(x)$

For minimality, suppose s is sufficient so that $p_\pi(\theta \mid x) = p_\pi(\theta \mid s(x))$, by assumption there exists a unique $\lambda : q_\lambda(\theta) = p(\theta \mid s)$ so $f_\pi : s \mapsto \lambda$ is well defined. □

Lemma 2 (Information theoretic equivalents)**.** *Suppose $\mathcal{P}$ is a convex set TFAE:*

1. *t is a sufficient statistic*
2. *$I[X; \theta] = I[T; \theta]$ for all π*
3. *$I[X; \theta \mid T] = 0$ for all π*

and t is minimal sufficient iff $I[X; T] = \min_{t' \in \mathcal{S}} I[X; T']$, where $\mathcal{S}$ is the class of sufficient statistics.

Proof. For sufficiency 1. $\iff$ 2.: Kullback & Leibler (1951) We have $D_{KL}[p(x \mid \theta_1) \mid\mid p(x \mid \theta_2)] = D_{KL}[p(t \mid \theta_1) \mid\mid p(t \mid \theta_2)]$ for all θ_1, θ_2 iff t is sufficient. Then note $p_\pi(x) = \int p(x \mid \theta)d\pi \in \mathcal{P}$ by convexity so there exists θ' such that $p_\pi(x) = p(x \mid \theta')$. This means $I[X; \theta] = \mathbb{E}_\pi[D_{KL}[p(x \mid \theta) \mid\mid p(x \mid \theta')]] = \mathbb{E}_\pi[D_{KL}[p(t \mid \theta) \mid\mid p(t \mid \theta')]]$ z t is sufficient.

2 $\iff$ 3. is an application of the d.p.i. $I[X; \theta \mid T] = I[\theta; (T, X)] - I[\theta; T] = I[X; T] - I[X; \theta]$

The minimality statement, due to Shamir et al. (2010).

□

Lemma 3 (IB, Shamir et al. (2010))**.** *Suppose*

$$t(\beta) \in \arg\min_{t \in \mathcal{F}} I[X; T] - \beta I[\theta; T] \tag{3}$$

and the class $\mathcal{F}$ is large enough, there exists β_c s.t. $t(\beta)$ is minimal sufficient on $supp(\pi)$ for any $\beta \geq \beta_c$.

A.2 Multiview Sufficiency

For the multiview environment we suppose we have an underlying data distribution $\pi(\theta)$ and a markov kernel (produced by augmentations) $l(x \mid \theta) = \int p(x \mid s, \theta)p(s)ds$. On a single draw $\theta' \sim \pi(\theta)$ we assume access to a set of views $x_{1:n} \sim l(x \mid \cdot)$

Assumption 3. *For any $\theta_1, \theta_2 \in \Theta$ there exists $x, x' \in X$ such that the likelihood ratio's are not equal:*

$$\frac{l(x' \mid \theta_1)}{l(x' \mid \theta_2)} \neq \frac{l(x \mid \theta_1)}{l(x \mid \theta_2)} \tag{4}$$

Or in multiview terms:

$$\frac{\mathbb{E}_{p_s}[f(x \mid \theta_1, s)]}{\mathbb{E}_{p_s}[f(x \mid \theta_2, s)]} \neq \frac{\mathbb{E}_{p_s}[f(x' \mid \theta_1, s)]}{\mathbb{E}_{p_s}[f(x' \mid \theta_2, s)]} \tag{5}$$

Definition A.3 (Multiview Sufficiency)**.** *We call $t : X \to Y$ multiview sufficient for (X, X') if for any $\pi(\theta)$ the induced distribution $p_\pi(x, x') = \int l(x \mid \theta)l(x' \mid \theta)\pi(\theta)d\theta$ factorises $p_\pi(x, x') = f(t(x), x')h(x)$ and minimal if for any other multiview sufficient s, $\exists f$ such that $f(s) = t$.*

Lemma 4 (Multiview sufficiency)**.** *Under Assumption 3. If t is multiview sufficient then it is also sufficient for θ.*

Proof. First note for dominated sets of measures pairwise sufficiency implies sufficiency Halmos & Savage (1949). So it is sufficient to consider whether t is sufficient for $l(x \mid \theta_1), l(x \mid \theta_2)$ and hence any discrete prior over this set. Then since t is sufficient for x' we have:

$$\sum_{i=1,2} [p(\theta_i \mid x) - p(\theta_i \mid t)]l(x' \mid \theta_i) = 0 \tag{6}$$

Let $\epsilon(\theta_i) = p(\theta_i \mid x) - p(\theta_i \mid t)$, then Equation (6) implies either $\epsilon(\theta_i) = 0$ for $i = 1, 2$ or:

$$-\frac{\epsilon(\theta_1)}{\epsilon(\theta_2)} = \frac{l(x' \mid \theta_1)}{l(x' \mid \theta_2)} \quad \forall x' \tag{7}$$

Since the l.h.s is independent of x' we can rule this out by Assumption 3, hence $\epsilon(\theta_i) = 0$ and $p(\theta_i \mid x) = p(\theta_i \mid t)$.

□

Lemma 5 (Minimal multiview sufficiency)**.** *If t is multiview minimal sufficient for (X, X') and Assumption 3 holds then t is minimal sufficient for θ*

Proof. First note if t is sufficient for θ then it is also multiview sufficient. Since $X - T - \theta - X'$ forms a markov chain. Now if S is an (X, θ) sufficient statistic then it is an (X, X') sufficient statistic by Lemma 4, so there exists $T = f(S)$, hence T is minimal. □

Proposition 3 (Multiview IB)**.** *Suppose Assumption 3 holds and (given a distribution π),*

$$t(\beta) \in \arg\min_{t \in \mathcal{F}} I[T; X \mid X'] - \beta I[X'; T] \tag{8}$$

there exists β_c s.t. $t(\beta)$ is minimal sufficient on $supp(\pi)$ for any $\beta > \beta_c$.

Proof. By Lemma 5 we just need to show minimal sufficiency for (X, X') however this holds by first noting $I[T; X] = I[T; X \mid X'] - I[X'; T]$, so for fixed $I[X'; T]$ minimising $I[T; X]$ is equivalent to minimising $I[T; X \mid X']$ so we can apply the logic of Lemma 3. □

Corollary 1 (Multiview Posteriors)**.** *Suppose the conditions of Proposition 3 with $\beta > \beta_c$ and a statisitical manifold $Q = \{q_\lambda(\theta) : \lambda \in \Lambda\}$ there exists a function $f_\pi : T \to \Lambda$ such that $p(\theta \mid x) = q_{f_\pi(t_\beta(x))}(\theta)$.*

Proof. By Proposition 3 t_β is sufficient for (X, θ), and by Lemma 1, the map ψ to the parameter manifold is a minimal sufficient statistic, so there exists $\psi = f \circ t_\beta$ □

A.3 Other Functionals

A.3.1 VicReg

Lemma 6 (Invariance Term)**.** $\mathbb{E}[\frac{1}{2}\left(t(X) - t(X')\right)^2 \mid \theta] = Var[T \mid \theta]$ *and hence* $\mathbb{E}[\frac{1}{2}\left(t(X) - t(X')\right)^2] = \mathbb{E}_\pi[Var[T \mid \theta]]$

Proof. Let $\mu = \mathbb{E}[T(X) \mid \theta$ then:

$$\begin{aligned}
\mathbb{E}[(t(X) - t(X'))^2 \mid \theta] &= \mathbb{E}[(t(X) - \mu + \mu - t(X'))^2 \mid \theta] && (9)\\
&= \mathbb{E}[(t(X) - \mu)^2 + (t(X') - \mu)^2 + (t(X) - \mu)^T(\mu - t(X')) \mid \theta] && (10)\\
&= 2Var[T \mid \theta] + \underbrace{\mathbb{E}[\mathbb{E}[(t(x) - \mu)^T(\mu - t(X')) \mid X = x] \mid \theta]}_{0} && (11)
\end{aligned}$$

□

(Intuition) Note for conditionally gaussian random variables (T, θ) we have $\mathbb{E}_\pi[\ln Var[T \mid \theta]] = H[T \mid \theta] + C$. Next supposing the regulariser can be satisfied exactly $\mathbb{E}[t(X)t(X)^T] = I$, and that T is also marginally gaussian then we have $H[T] = C$ and $I[T; \theta] \propto -\mathbb{E}_\pi[\ln Var[T \mid \theta]] \leq -\ln \mathbb{E}_\pi[Var[T \mid \theta]]$ and so under these assumptions we are in the infomax case, see Zbontar et al. (2021).

A.3.2 INFONCE

Lemma 7. *Suppose Assumption 3 holds and for some π,*

$$t \in \arg\min_{t \in \mathcal{F}} -\mathbb{E}_p[t(X)^t t(X')] + \ln \mathbb{E}_q[e^{-t(X)^t t(X')}] \tag{12}$$

Then t is sufficient for θ on $supp(\pi)$.

Proof.

$$L(t) = \mathbb{E}_{X'}\left[\underbrace{-\mathbb{E}_p[t(X)^t t(x') \mid X' = x'] + \ln \mathbb{E}_q[e^{-t(X)^t t(x')} \mid X' = x']}_{f(t,X')}\right] \tag{13}$$

then $-f(t, X') \leq D_{KL}[P_{x'} \mid\mid Q]$ and $-L(t) \leq I[X; X']$ where $P_{x'}(dx) = p(x \mid x')dx$ and $Q(dx) = p(x)$ by Donsker-Varadhan representation. Further equality is achieved iff $\frac{p(x|x')}{p(x)} = \frac{e^{t(x)^t t(x')}}{Z}$ where $Z = \mathbb{E}_p[e^{t(x)^t t(x')}]$ therefore $p(x \mid x') = p(x)f(t(x), t(x'))$ which is the factorisation condition for sufficiency on (X, X') then applying Lemma 4. □

Corollary 2 (InfoNCE Posteriors)**.** *Suppose the conditions of Proposition 3 there exists a function $f_\pi : T \to \Lambda$ such that $p(\theta \mid x) = q_{f_\pi(t_\beta(x))}(\theta)$.*

Proof. By Lemma 7 t_β is sufficient for (X, θ), and by Lemma 1, the map ψ to the parameter manifold is a minimal sufficient statistic, so there exists $\psi = f \circ t_\beta$ □

B EXPERIMENT DETAILS

For the first task, we define three types of f_{MLP}: identity, fully connected layer (10,10), and two fully connected layers (10,10) with an intermediate nonlinear activation function. For the InfoNCE encoder, we chose between 2 different architectures depending on whether the MLP includes a nonlinearity. In the identity or linear case the architecture is defined as: Dense(10, 1024), Dense(1024, 10). For the nonlinear scenario we do the following: Dense(10, 1024), nonlinearity, Dense(1024, 1024), nonlinearity, Dense(1024, 1024), nonlinearity, Dense(1024, 10).

For the second task, we collapse the second MLP weight matrix by creating a mapping from 10 to 5 dimensions. The schematic architecture here is the following: Dense(10, 10) - leaky relu - Dense(10, 5). The encoder here uses a leaky relu and has the same structure as in the first task.

Across all experiments we trained the SSL for 20k steps, with a learning rate of 0.0001 using Adam optimizer. To train the SVI model we rely on the SVI class of Numpyro and we use build in loss function $Trace\_ELBO$. We set the learning rate to 0.0005 using the Adam optimizer. Then we do 10k inference steps on a data point per data point basis for over 1k datapoints. Similarly, we get the latent representations of the SSL for that same data to compute the R^2 score between each approach.

To make sure extra latent dimensions learn to capture uncertainty and it's not due to a byproduct of increasing latent size, we measure the R^2 for an untrained InfoNCE network. As can be seen in Figure 5, the SSL encodings obtained with random weights is significantly lower than the one we report in Figure 4.

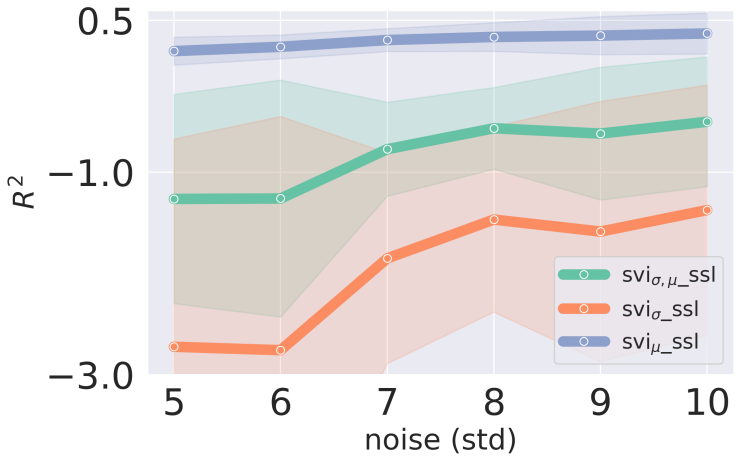


Figure 5: Baseline comparison for the second toy model.