Likelihood Ratio Exponential Families

Anonymous Author(s)
Affiliation
Address
email

Abstract

The exponential family is well known in machine learning and statistical physics as the maximum entropy distribution subject to a set of observed constraints [1], while the geometric mixture path is common in MCMC methods such as annealed importance sampling (AIS) [2, 3]. Linking these two ideas, Brekelmans et al. [4] interpret the geometric mixture path as an exponential family of distributions to analyse the recent thermodynamic variational objective (TVO) [5].

In this work, we extend *likelihood ratio exponential families* to include solutions to rate-distortion (RD) optimization [6, 7], the Information Bottleneck method (IB) method [8], and recent rate-distortion-classification (RDC) approaches combining RD and IB [9, 10]. We provide a common mathematical framework for understanding these methods using the conjugate duality of exponential families. Further, we collect existing results [11–13] to express intermediate distributions via a variational representation related to hypothesis testing and the Neyman Pearson lemma [14, 15], and leverage this perspective to identify the point at which the TVO integrand, or expected likelihood ratio, matches the log partition function.

1 Introduction

Likelihood Ratio Exponential Family Following Brekelmans et al. [4], we consider the geometric mixture path between a base distribution $\pi_0(z)$ and target $\pi_1(z)$ or posterior $\pi_1(z|x)$, as an exponential family. We define the sufficient statistics $\phi(z) = \log \pi_1(z)/\pi_0(z)$ as the log likelihood ratio [4], although in practice it is convenient to consider an unnormalized target $\pi_1(z) \propto \tilde{\pi}_1(z)$ or $\pi_1(z|x) \propto \tilde{\pi}_1(x,z)$ and adjust the normalization constant accordingly. Using a natural parameter β and base distribution π_0,

$$\pi_\beta(z) = \pi_0(z) \exp\{\beta \cdot \phi(z) \; - \psi(\beta)\} = \frac{1}{Z_\beta} \pi_0(z)^{1-\beta} \tilde{\pi}_1(z)^\beta \tag{1}$$

$$\text{where} \;\; \phi(z) := \log \frac{\tilde{\pi}_1(z)}{\pi_0(z)} \qquad \psi(\beta) := \log Z_\beta = \log \int \pi_0(z)^{1-\beta} \tilde{\pi}_1(z)^\beta dz \tag{2}$$

Before discussing examples in Sec. 2, we review background on conjugate duality in exponential families, which provides insights which are not evident from writing (1) as a geometric mixture [4].

Legendre Duality in Exponential Families Since the log partition function $\psi(\beta)$ of an exponential family is convex in the natural parameters β, its gradient will be unique and may be used as a dual parameterization for π_β [16, 17]. This diffeomorphism between the natural parameters $\beta = \{\beta_j\}$ [1] and moment parameters, denoted $\eta = \{\eta_j\}$, also defines the convex conjugate function $\psi^*(\eta)$, with

$$\psi^*(\eta) = \sup_\beta \beta \cdot \eta - \psi(\beta) \qquad \Longrightarrow \; \eta_j = \frac{\partial \psi}{\partial \beta} = \mathbb{E}_{\pi_\beta}[\phi_j(x,z)] \; \forall \; j \tag{3}$$

[1] We allow for multiple sufficient statistics, with $\beta \cdot \phi(z) = \sum_j \beta_j \cdot \phi_j(z)$ denoting the dot product.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

With the Lebesgue or counting measure as $\pi_0(z)$, the conjugate $\psi^*(\eta)$ corresponds to the negative entropy of the maximum entropy solution $\pi_\beta(z)$ with observable constraint η [18, 17]. With a general base measure (e.g. [4] App. A), we have

$$\psi^*(\eta_\beta) = D_{KL}[\pi_\beta(z|x)||\pi_0(z)] \tag{4}$$

Since the convex conjugate is an involution, $(\psi^*)^* = \psi$, we can obtain a similar optimization to (3) in terms of $\psi(\beta) = \sup_\eta \beta \cdot \eta - \psi^*(\eta)$. This leads to the canonical expression for Legendre duality, when the two optimizations are in equilibrium and the vectors η_β and β are in correspondence [18]

$$\psi^*(\eta_\beta) + \psi(\beta) - \beta \cdot \eta_\beta = 0\,. \tag{5}$$

Finally, we can construct Bregman divergences from the convex functions $\psi(\beta)$ or $\psi^*(\eta)$. Using (2) and (5), $D_\psi[\beta : \beta'] := \psi(\beta) - \psi(\beta') - \langle \beta - \beta', \nabla\psi(\beta')\rangle = D_{\psi^*}[\eta_{\beta'} : \eta_\beta] = D_{KL}[\pi_{\beta'}||\pi_\beta]$ [16].

2 Examples

Thermodynamic Variational Objective In the variational autoencoder (VAE) setting, the TVO [5, 4] uses the approximate posterior as the initial distribution $\pi_0 = q(z|x)$ and joint generative model as the unnormalized target $\tilde{\pi}_1 = p_\theta(x, z)$. Masrani et al. [5] use thermodynamic integration (TI) [19, 20] to express $\psi(x; 1) = \log Z_1(x) = \log p_\theta(x)$ as an integral over the geometric path (2),

$$\log Z_1(x) - \log Z_0(x) = \int_0^1 \frac{d}{d\beta} \log Z_\beta \, d\beta = \int_0^1 \mathbb{E}_{\pi_\beta}\big[\phi(z)\big]\, d\beta\,. \tag{6}$$

where we use the fact that the (partial) derivative of the log partition function equals the expected sufficient statistics in any exponential family [16, 17]. Since $\psi(x; \beta)$ is convex in β for any x, the left- and right-Riemann sums will provide lower and upper bounds on the log marginal likelihood,

$$\sum_{t=0}^{T-1} (\beta_{t+1} - \beta_t) \cdot \mathbb{E}_{\pi_{\beta_t}}\big[\log \frac{\tilde{\pi}_1(x,z)}{\pi_0(z)}\big] \;\le\; \log Z_1 \;\le\; \sum_{t=1}^{T} (\beta_t - \beta_{t-1}) \cdot \mathbb{E}_{\pi_{\beta_t}}\big[\log \frac{\tilde{\pi}_1(x,z)}{\pi_0(z)}\big]\,. \tag{7}$$

We derive novel insights on TVO curve via hypothesis testing in Sec. 3. Note that TI bounds as in (7) may be constructed for any one-dimensional likelihood ratio exponential family, such as in RD, although more care would be required for multiple sufficient statistics as in RDC below [9, 10].

Rate-Distortion Rate-distortion (RD) optimization ([6, 8, 21, 22, 7] Ch. 13) formalizes the problem of lossy compression subject to a fidelity constraint. As in Alemi et al. [6][10], we measure the rate using the KL divergence to a fixed marginal distribution $\pi_0(z) = m(z)$, which upper bounds the mutual information in general. The distortion function $d(x, z)$ measures the quality of a code z. RD optimization seeks the minimum-rate encoding which achieves a desired average distortion D,

$$R(D) = \min_{q(z|x)} D_{KL}[q(z|x)||m(z)] \quad \text{subj. to} \quad \mathbb{E}_{q(z|x)}[d(x|z)] \le D\,. \tag{8}$$

We restrict our attention to a reconstruction loss distortion $d(x, z) = -\log p_\theta(x|z)$ as in [6]. Introducing β to enforce the constraint, we obtain the unconstrained Lagrangian

$$\max_\beta \min_q D_{KL}[q(z|x)||m(z)] - \beta\big(\mathbb{E}_{q(z|x)}[d(x,z)] - D\big) \tag{9}$$

whose solution, for a given $m(z)$, has an exponential family form with $\phi(x, z) = -d(x, z)$ (e.g. [8])

$$q^*(z|x) = \frac{1}{Z_\beta(x)} m(z) \, \exp\{-\beta \cdot d(x,z)\} = \frac{1}{Z_\beta(x)} m(z)\, p_\theta(x|z)^\beta \tag{10}$$

From the likelihood ratio perspective, we can choose $\pi_0(z) = m(z)$ and $\tilde{\pi}_1(x, z) = p_\theta(x|z)m(z) \propto p_\theta(z|x)$. Absorbing the factor of $p_\theta(x)$ into the normalizer $Z_\beta(x)$, we obtain the sufficient statistics

$$\phi(x,z) = \log \frac{\tilde{\pi}_1(x,z)}{\pi_0(z)} = \log \frac{p_\theta(x|z)m(z)}{m(z)} = \log p_\theta(x|z) = -d(x,z)\,, \tag{11}$$

so that the solution $q^*(z|x)$ in (10) matches $\pi_\beta(z|x)$ in the likelihood ratio family induced by (11). The Lagrange multiplier β is chosen to enforce the distortion constraint D, which, since $\phi(x, z) = -d(x, z)$, translates to seeking β such that the moment parameters $\eta_\beta = -D$. At this optimal solution, $R(D)$ simply matches the conjugate $\psi^*(\eta)$ in (33)

$$R(D) = \psi^*(\eta) = D_{KL}[\pi_\beta(z|x)||m(z)] = \beta \cdot \eta - \psi(\beta) = -\beta\, D - \log Z_\beta(x)\,. \tag{12}$$

Huang et al. [22] use the expression in (12) to estimate the RD curve using AIS [2].

Information Bottleneck and RDC When defining 'relevant information' via a random variable such as a label y, the Information Bottleneck (IB) method [8, 23, 24] simplifies to an RD problem with a learned classifier providing the distortion function $c(y,z) = -\log p_\theta(y|z)$ ([8] or App.B).

$$\min_{q(z|x)} D_{KL}[q(z|x)||m(z)] \quad \text{subj. to} \quad \mathbb{E}_{q(z|x)}[c(y,z)] \leq C \tag{13}$$

Recent work [9, 10] considers 'RDC' optimization using both reconstruction and classification loss,

$$\min_{q(z|x)} D_{KL}[q(z|x)||m(z)] \quad \text{subj. to} \quad \mathbb{E}_{q(z|x)}[d(x,z)] \leq D \ , \ \mathbb{E}_{q(z|x)}[c(y,z)] \leq C \tag{14}$$

In this case, we may consider two sufficient statistics in our likelihood ratio exponential family. Similarly to multivariate IB [25, 26], we use an unnormalized target which factorizes as $\tilde{\pi}_1(x,y,z) = p_\theta(x|z)p_\theta(y|z)m(z)$, and consider the likelihood ratio sufficient statistics

$$\phi_d(x,z) = \log\frac{\pi_1(z|x)}{\pi_0(z)} = \log\frac{p_\theta(x|z)}{p_\theta(x)} \propto -d(x,z) \qquad \phi_c(y,z) = \log\frac{\pi_1(z|y)}{\pi_0(z)} = \log\frac{p_\theta(y,z)}{p(y)} \propto \log p_\theta(y|z) = -c(y,z) \tag{15}$$

where we again absorb $p_\theta(x)$ and $p(y)$ into the normalization. Introducing Lagrange multipliers $\beta = \{\beta_D, \beta_C\}$ to enforce $\eta_d(\beta) = -D$, $\eta_c(\beta) = -C$ at optimality, we obtain the solution of (14) as a geometric mixture [9, 10] belonging to the likelihood ratio family with $\phi = \{\phi_d, \phi_c\}$

$$\begin{aligned}\pi_\beta(z|x,y) &= m(z)\exp\big\{\beta_D \cdot \phi_d(x,z) + \beta_C \cdot \phi_c(y,z) - \psi(x,y;\beta)\big\} \\ &= \frac{1}{Z_\beta(x,y)}\, m(z)\, p_\theta(x|z)^{\beta_D}\, p_\theta(y|z)^{\beta_C}\end{aligned} \tag{16}$$

With applications in transfer learning, Gao and Chaudhari [9] seek to evolve model parameters θ and the approximate posterior $q(z|x)$ along an 'equilibrium surface' of optimal solutions to (14). We interpret their free energy $F(\beta_D,\beta_C)$, where β_D, β_C are analogous to the *intensive* variables of a physical system [10], as the negative log partition function $-\psi(\beta_D,\beta_C)$. Written using the conjugate optimization (3), we seek θ, $q(z|x)$ yielding the appropriate distortion and classification loss η_D, η_C

$$-F(\beta_D,\beta_C) = \psi(\beta_D,\beta_C) = \sup_{\eta_d,\eta_c} \beta_D\,\eta_d + \beta_C\,\eta_c - \psi^*(\eta_d,\eta_c) \tag{17}$$

Similarly, for given *extensive* variables η_D, η_C, the optimal rate $R(D,C)$ corresponds to $\psi^*(\eta_D,\eta_C)$

$$R(D,C) = \psi^*(\eta_D,\eta_C) = \sup_{\beta_d,\beta_c} -\beta_d\,D - \beta_c\,C - \psi(\beta_d,\beta_c)\,, \tag{18}$$

At optimality on the 'equilibrium surface' [9], we have $q(z|x) = \pi_\beta(z|x)$, which fulfills the constraints $\eta_\beta = \{\eta_D,\eta_C\} = \{-D,-C\}$ for $\beta = \{\beta_D,\beta_C\}$ and the current decoder and classifier parameters θ. This corresponds to equality in the canonical Legendre duality equation (5)

$$\psi^*(\eta_D,\eta_C) + \psi(\beta_D,\beta_C) - \beta_D\,\eta_D - \beta_C\,\eta_C = 0\,. \tag{19}$$

and leads to the 'first law of learning' from [10] when $\psi(\beta_D,\beta_C)$ is considered as a fixed quantity.

3 Variational Representations and Hypothesis Testing

Grosse et al. [11] note that any distribution along the geometric mixture path can be given a variational representation as the solution to an expected KL divergence minimization

$$\pi_{\beta_t}(z) = \arg\min_{r(z)} (1-t)\,D_{KL}[r(z)||\pi_{\beta_0}(z)] + \,t\,D_{KL}[r(z)||\pi_{\beta_1}(z)] \tag{20}$$

In this section, we intepret (20) as a Bregman information (or gap in Jensen's inequality) [12], or as describing an optimal decision rule for hypothesis testing using the Neyman Pearson lemma.

Bregman Information Banerjee et al. [12] define the *Bregman information* as the minimum expected divergence to a representative point in the second argument. Regardless of the divergence considered, the optimal representative corresponds to the mean over the arguments. Since $D_{KL}[r(z)||\pi_{\beta_0}(z)] = D_\psi[\beta_0 : \beta_r]$ for $r(z)$ within the exponential family, we can rewrite (20) as

$$\beta_t = \arg\min_{\beta_r} (1-t)\,D_\psi[\beta_0:\beta_r] + \,t\,D_\psi[\beta_1:\beta_r] \;\; \text{where} \;\; \beta_t = (1-t)\cdot\beta_0 + t\cdot\beta_1 \tag{21}$$

At this optimum, the expected KL divergence (21) can be written as a gap in Jensen's inequality for the convex function $\psi(\beta)$ [12], or, as shown in [27] or App. C, as a Rényi divergence with order t

$$\begin{aligned}(1-t)\,D_\psi[\beta_0:\beta_t] + \,t\,D_\psi[\beta_1:\beta_t] &= (1-t)\,\psi(\beta_0) + t\,\psi(\beta_1) - \psi\big(\beta_t\big) \\ &= (1-t)\,D_t[\pi_{\beta_1}:\pi_{\beta_0}]\end{aligned} \tag{22}$$

Neyman Pearson Lemma Suppose we have access to n i.i.d. observations from an unknown distribution $r(z)$, and are interested in testing the hypotheses that either $H_0 : r(z) = \pi_0(z)$ or $H_1 : r(z) = \pi_1(z)$. The Neyman-Pearson lemma states that the likelihood ratio test is optimal, in the sense that, for any other decision region with type-1 error $Pr(e_1) = R$, then the type-2 error is no better than that of the likelihood ratio test ([7] Ch. 11, [14]). The decision rule is given by

$$A_n(\pi_1; \eta) = \left\{ z_{1:n} \;\middle|\; \frac{1}{n} \sum_{i=1}^{n} \log \frac{\pi_1(z_i)}{\pi_0(z_i)} \geq \eta \right\} \tag{23}$$

for some threshold η. Let a type-1 error occur when n i.i.d. draws $\{z_i\}_{i=1}^N$ from $\pi_0(z)$ will yield empirical expectations exceeding the threshold η. Sanov's Theorem and large deviation theory ([7] Ch. 11, [28, 15]) states that the asymptotic error exponent corresponds to a KL divergence

$$\lim_{n\to\infty} \frac{1}{n} Pr(e_1) \to \exp\{-D_{KL}[r^*(z)||\pi_0(z)]\} \;\text{ where } r^*(z) = \min_{r(z)\in\mathcal{M}_\eta} D_{KL}[r(z)||\pi_0(z)] \tag{24}$$

and feasibile set $\mathcal{M}_\eta := \{r(z) \,|\, \mathbb{E}_r \log \frac{\pi_1(z)}{\pi_0(z)} = \eta\}$ reflects a moment constraint. With $\psi^*(\eta) = D_{KL}[\pi_{\beta_\eta}(z)||\pi_0(z)]$ as in (33), this corresponds exactly to the conjugate or maximum entropy optimization for a given expected likelihood ratio threshold, and thus $r^*(z)$ lies within our exponential family,

$$r^*(z) = \pi_0(z) \exp\{\beta_\eta \cdot \log \frac{\pi_1(z)}{\pi_0(z)} - \psi(\beta)\} \tag{25}$$

As shown in Fig. 1, Sanov's Theorem implies a similar expression for the asymptotic type-2 error, when draws from $\pi_1(z)$ achieve a *lower* expected likelihood ratio than η. Expressing the conditions of the Neyman Pearson lemma using these asymptotic error probabilities [2] , we can write

$$Pr(e_2) = \min_{r(z)} D_{KL}[r(z)||\pi_1(z)] \quad \text{subj. to} \quad D_{KL}[r(z)||\pi_0(z)] = R \tag{26}$$

Using a Lagrange multiplier $\lambda = \frac{1-\beta}{\beta}$ to enforce the constraint, we obtain the variational form (20)

$$\frac{1}{\beta} Pr(e_2) = \min_{r(z)} (1-\beta) D_{KL}[r(z)||\pi_0(z)] + \beta D_{KL}[r(z)||\pi_1(z)] \tag{27}$$

Thus, any distribution in our likelihood ratio exponential family corresponds to a likelihood ratio test with decision threshold η, which is optimal for a type-1 error region of size $\psi^*(\eta) = R$.

Chernoff Information While each choice of β_η determines a likelihood ratio test and error region, how should we choose this parameter? Regardless of the prior probabilities p_0, p_1 which we might assign to each hypothesis in a Bayesian setting, the Chernoff information provides the best achievable error exponent in the large sample limit ([13], [7] Ch. 11).

$$C^* = -\min_\beta \log \int \pi_0(z)^{1-\beta} \pi_1(z)^\beta dz = \max_\beta \, (1-\beta)\, \psi(0) + \beta\, \psi(1) - \psi\big(\beta\big) \tag{28}$$

At this optimum, denoted the Chernoff point [13], we show in Appendix D that

$$D_{KL}[\pi_{\beta^*}(z)||\pi_0(z)] = D_{KL}[\pi_{\beta^*}(z)||\pi_1(z)] \tag{29}$$

and the optimal decision rule is given by a threshold of $\eta_{\beta^*} = \mathbb{E}_{\pi_{\beta^*}} \log \frac{\pi_1(z)}{\pi_0(z)} = 0$.

Chernoff Point on the TVO Integrand For the unnormalized likelihood ratio $\log \tilde{\pi}_1(z)/\pi_0(z)$, we can intepret the Chernoff point using thermodynamic integration bounds (7)

$$\sum_{t=0}^{T-1} (\beta_{t+1} - \beta_t) \cdot \mathbb{E}_{\pi_{\beta_t}} \big[\log \frac{\tilde{\pi}_1(x,z)}{\pi_0(z)} \big] \;\leq\; \log Z_1 \;\leq\; \sum_{t=1}^{T} (\beta_t - \beta_{t-1}) \cdot \mathbb{E}_{\pi_{\beta_t}} \big[\log \frac{\tilde{\pi}_1(x,z)}{\pi_0(z)} \big], \tag{30}$$

With $\pi_0(z) = q(z|x)$ as in TVO [5, 4], we note that the integrand at $\beta = 0$ corresponds to the familiar evidence lower bound (ELBO), $\mathbb{E}_{\pi_0} \big[\log \frac{\tilde{\pi}_1(x,z)}{\pi_0(z)} \big] = \log Z_1(x) - D_{KL}[\pi_0(z)||\pi_1(z|x)]$. Similarly, at $\beta = 1$, the integrand $\mathbb{E}_{\pi_1}[\cdot] = \log Z_1(x) + D_{KL}[\pi_1(z|x)||\pi_0(z)]$ provides an upper bound. The Chernoff point determines where the moment parameters switch from an lower bound to an upper bound, or β^* such that $\eta_{\beta^*} = \mathbb{E}_{\pi_{\beta^*}}[\cdot] = \log p_\theta(x)$. We visualize this in Fig. 2, noting that the shaded regions corresponding to the KL divergence (see [4]) will have equal area due to (29).

[2] While Neyman-Pearson is often obtained via the method of types [7], Csiszár [29] treat the continuous case.

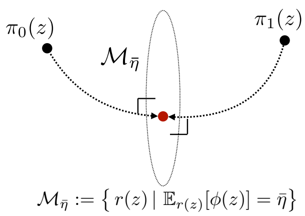

Figure 1: Sanov's Theorem. (See [18] Sec. 1.6, 2.8 for an interpretation in terms of projection and a generalization of the Pythagorean Theorem)

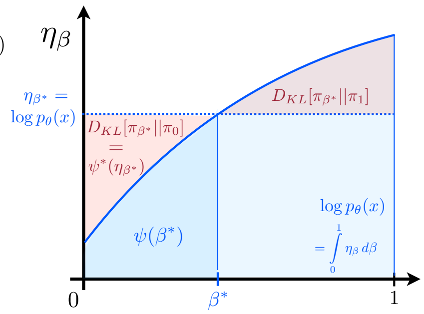

Figure 2: Chernoff point on $\eta_\beta = \nabla\psi(\beta)$.

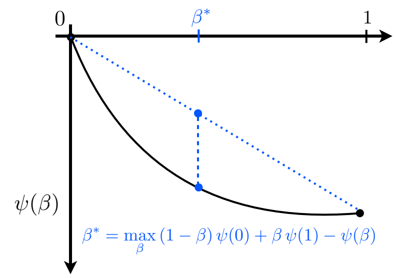

Figure 3: Chernoff point on $\psi(\beta)$

References

[1] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[2] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

[3] Christopher Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56(5):5018, 1997.

[4] Rob Brekelmans, Vaden Masrani, Frank Wood, Greg Ver Steeg, and Aram Galstyan. All in the exponential family: Bregman duality in thermodynamic variational inference. *International Conference on Machine Learning*, 2020.

[5] Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. *Advances in Neural Information Processing Systems*, 2019.

[6] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

[7] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[8] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pages 368–377, 1999.

[9] Yansong Gao and Pratik Chaudhari. A free-energy principle for representation learning. *International Conference on Machine Learning*, 2020.

[10] Alexander A Alemi and Ian Fischer. Therml: Thermodynamics of machine learning. *arXiv preprint arXiv:1807.04162*, 2018.

[11] Roger B Grosse, Chris J Maddison, and Ruslan R Salakhutdinov. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems*, pages 2769–2777, 2013.

[12] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[13] Frank Nielsen. An information-geometric characterization of Chernoff information. *IEEE Signal Processing Letters*, 20(3):269–272, 2013.

[14] Shashi Borade and Lizhong Zheng. I-projection and the geometry of error exponents. In *in Allerton Conference*. Citeseer, 2006.

[15] Imre Csiszár and Paul C Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[16] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[17] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[18] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[19] Yosihiko Ogata. A monte carlo method for high dimensional integration. *Numerische Mathematik*, 55(2):137–157, 1989.

[20] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

[21] Kenneth Rose, Eitan Gurewitz, and Geoffrey Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594, 1990.

[22] Sicong Huang, Alireza Makhzani, Yanshuai Cao, and Roger Grosse. Evaluating lossy compression rates of deep generative models. In *International Conference on Machine Learning*.

[23] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897–2905, 2018.

[24] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[25] Noam Slonim, Nir Friedman, and Naftali Tishby. Multivariate information bottleneck. *Neural computation*, 18(8):1739–1789, 2006. ISSN 0899-7667. doi: 10.1162/neco.2006.18.8.1739.

[26] Gal Elidan and Nir Friedman. The information bottleneck em algorithm. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 200–208, 2002.

[27] Frank Nielsen and Richard Nock. On Rényi and Tsallis entropies and divergences for exponential families. *arXiv preprint arXiv:1105.3259*, 2011.

[28] Imre Csiszár. Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

[29] Imre Csiszár. The method of types [information theory]. *IEEE Transactions on Information Theory*, 44(6):2505–2523, 1998.

A Conjugate as a KL Divergence

When considering an exponential family of the form

$$\pi_\beta(z) = \pi_0(z) \exp\{\beta \cdot \phi(z) \, - \psi(\beta)\} \, . \tag{31}$$

we show that $\psi^*(\eta)$ takes the form of a KL divergence when considering a base measure $\pi_0(z)$.

$$\begin{aligned}
\psi^*(\eta) &= \sup_\beta \beta \cdot \eta - \psi(\beta) &(32)\\
&= \beta_\eta \cdot \eta - \psi(\beta_\eta) \\
&= \mathbb{E}_{\pi_{\beta_\eta}}[\beta_\eta \cdot \phi(z)] - \psi(\beta_\eta) \\
&= \mathbb{E}_{\pi_{\beta_\eta}}[\beta_\eta \cdot \phi(z)] - \psi(\beta_\eta) \pm \mathbb{E}_{\pi_{\beta_\eta}}[\log \pi_0(z)] \\
&= \mathbb{E}_{\pi_{\beta_\eta}}[\log \pi_{\beta_\eta(z)} - \log \pi_0(z)] \\
&= D_{KL}[\pi_{\beta_\eta}(z) || \, \pi_0(z)] &(33)
\end{aligned}$$

where we have added and subtracted a factor of $\mathbb{E}_{\pi_{\beta_\eta}} \log \pi_0(z)$ in the fourth line. When $\pi_0(z)$ is constant with respect to z, $D_{KL}[\pi_{\beta_\eta}(z) || \, \pi_0(z)]$ reduces to the familiar definition of the conjugate function ψ^* as the negative entropy $\mathbb{E}_{\pi_{\beta_\eta}} \log \pi_{\beta_\eta}(z)$ [17].

B Information Bottleneck as Rate-Distortion

The Information Bottleneck (IB) method [8] defines the 'relevant information' in a representation, $I(Y:Z)$, via another variable of interest Y, often taken to be a label. The IB objective then seeks a minimal encoding Z which maintains a given level of predictive ability about the target.

$$\min_{q(z|x)} \;\; I_q(X;Z) \;\; \text{subj. to. } \;\; I_q(Y;Z) \geq I_c \tag{34}$$

where we let I_q reflect the exact mutual information for the true data and label distributions $q(x)q(y|x)$ with a given encoding function $q(z|x)$.

When the desired information constraint equals the total information $I_c = I_q(X;Y)$ that the data source contains about the label, (34) corresponds to the problem of finding the minimal sufficient statistics z for y with respect to x. The IB objective generalizes this optimization for smaller values of I_c.

Since $I_q(Y;Z) = H_q(Y) - H_q(Y|Z) = -\mathbb{E}_q \log q(y) + \mathbb{E}_q \log q(y|z)$, we can ignore the label entropy as a constant with respect to z. While it may be difficult to obtain the true posterior $q(y|z)$ of the labels given latent variables , we can instead optimize a variational classifier $p(y|z)$. This provides an lower bound on the mutual information since $D_{KL}[q(y|z)||p(y|z)] \geq 0$ and is also known as the 'test channel' in rate-distortion theory ([7] Ch. 13). Applying this inequality within the unconstrained IB Lagrangian,

$$\begin{aligned}
\mathcal{L}_{IB} &= \max_\beta \min_{q(z|x)} \;\; I_q(X;Z) - \beta \big(- \mathbb{E}_q \log q(y) + \mathbb{E}_q \log q(y|z) - I_c \big) \\
&\geq \max_\beta \min_{q(z|x), p(y|z)} \;\; I_q(X;Z) - \beta \big(- \mathbb{E}_q \log q(y) + \mathbb{E}_q \log p(y|z) - I_c \big) \\
&= \max_\beta \min_{q(z|x), p(y|z)} \;\; I_q(X;Z) - \beta \, \mathbb{E}_{p(y(x),z)}[p(y|z)] + \text{const} &(35)
\end{aligned}$$

where $y(x)$ indicates the label of a given data point.

As shown in Tishby et al. [8], the Information Bottleneck is a special case of rate-distortion with

$$c(y(x),z) = D_{KL}[q(y|x)||q(y|z)] = \mathbb{E}_q[q(y|x)] - \mathbb{E}_q[q(y|z)] \tag{36}$$

Comparing (35) with (36), note that $\mathbb{E}_q[q(y|x)]$ is a constant, leaving $c(y(x),z) = -\mathbb{E}_{q(y(x)|z)}[q(y|z)]$ as the effective distortion measure. If this quantity is intractable, we can instead define the distortion function using $p(y|z)$ as above.

C Rényi Divergence as a Jensen Gap

We consider the Rényi α divergence between any two distributions π_{β_1} and π_{β_0} in our exponential family, so that $\pi_\beta(z|x) = \pi_0(z)^{1-\beta}\pi_1(z)^\beta / Z_\beta(x)$. Noting that the scaling factor $\alpha - 1 \leq 0$, we proceed to show that the scaled divergence is equal to a gap in Jensen's inequality:

$$
\begin{aligned}
&(1-\alpha)D_\alpha[\pi_{\beta_1}(z) : \pi_{\beta_0}(z)] \\
&\qquad = (1-\alpha)\frac{1}{\alpha-1}\log \int \pi_{\beta_0}^{1-\alpha}\pi_{\beta_1}^{\alpha}\, d\mu \\
&\qquad = -\log \int \big(\frac{\pi_0^{1-\beta_0}\pi_1^{\beta_0}}{Z_{\beta_0}}\big)^{1-\alpha}\big(\frac{\pi_0^{1-\beta_1}\pi_1^{\beta_1}}{Z_{\beta_1}}\big)^{\alpha} d\mu \\
&\qquad = -\bigg(\log \int \pi_0^{1-\beta_0-\alpha+\alpha\beta_0+\alpha-\alpha\beta_1}\pi_1^{\beta_0-\alpha\beta_0+\alpha\beta_1} d\mu - \big((1-\alpha)\log Z_{\beta_0} + \alpha \log Z_{\beta_1}\big)\bigg) \\
&\qquad = -\bigg(\log \int \pi_0^{1-[(1-\alpha)\beta_0+\alpha\beta_1]}\pi_1^{(1-\alpha)\beta_0+\alpha\beta_1} d\mu - \big((1-\alpha)\log Z_{\beta_0} + \alpha \log Z_{\beta_1}\big)\bigg) \\
&\qquad = (1-\alpha)\psi(\beta_0) + \alpha\psi(\beta_1) - \psi\big((1-\alpha)\beta_0 + \alpha\beta_1\big) \\
&\qquad = \mathcal{J}_{\alpha,\psi}
\end{aligned}
$$

D Equal KL Divergences Derivation

We show that the KL divergences that constitute $\mathcal{J}_{\alpha,\psi}$ are equal at the critical point $\eta_\alpha = \frac{\psi(\beta_1)-\psi(\beta_0)}{\beta_1-\beta_0}$:

$$
\begin{aligned}
D_\psi[\beta_0 : \beta_\alpha] &= \psi(\beta_0) - \psi(\beta_\alpha) - (\beta_0 - \beta_\alpha)\eta_\alpha \\
&= \psi(\beta_0) - \psi(\beta_\alpha) + \frac{(\beta_\alpha - \beta_0)}{\beta_1 - \beta_0}(\psi(\beta_1) - \psi(\beta_0)) \\
&= \frac{1}{\beta_1 - \beta_0}\Big((\beta_1 - \beta_0)\psi(\beta_0) - (\beta_1 - \beta_0)\psi(\beta_\alpha) + (\beta_\alpha - \beta_0)\psi(\beta_1) - (\beta_\alpha - \beta_0)\psi(\beta_0)\Big) \\
&= \frac{1}{\beta_1 - \beta_0}\Big((\beta_1 - \beta_\alpha)\psi(\beta_0) + (\beta_\alpha - \beta_0)\psi(\beta_1) - (\beta_1 - \beta_0)\psi(\beta_\alpha)\Big) \\
&= \Big(\frac{\beta_1 - \beta_\alpha}{\beta_1 - \beta_0}\psi(\beta_0) + \frac{\beta_\alpha - \beta_0}{\beta_1 - \beta_0}\psi(\beta_1) - \psi(\beta_\alpha)\Big)
\end{aligned}
$$

$$
\begin{aligned}
D_\psi[\beta_1 : \beta_\alpha] &= \psi(\beta_1) - \psi(\beta_\alpha) - (\beta_1 - \beta_\alpha)\eta_\alpha \\
&= \psi(\beta_1) - \psi(\beta_\alpha) - \frac{(\beta_1 - \beta_\alpha)}{\beta_1 - \beta_0}(\psi(\beta_1) - \psi(\beta_0)) \\
&= \frac{1}{\beta_1 - \beta_0}\Big((\beta_1 - \beta_0)\psi(\beta_1) - (\beta_1 - \beta_0)\psi(\beta_\alpha) - (\beta_1 - \beta_\alpha)\psi(\beta_1) + (\beta_1 - \beta_\alpha)\psi(\beta_0)\Big) \\
&= \frac{1}{\beta_1 - \beta_0}\Big((\beta_1 - \beta_\alpha)\psi(\beta_0) + (\beta_\alpha - \beta_0)\psi(\beta_1) - (\beta_1 - \beta_0)\psi(\beta_\alpha)\Big) \\
&= \Big(\frac{\beta_1 - \beta_\alpha}{\beta_1 - \beta_0}\psi(\beta_0) + \frac{\beta_\alpha - \beta_0}{\beta_1 - \beta_0}\psi(\beta_1) - \psi(\beta_\alpha)\Big)
\end{aligned}
$$

We have shown that the two divergences are equal when our condition on η_α holds. Further, observe that each divergence amounts to a Jensen gap $\mathcal{J}_{\alpha,\psi}$ with $\alpha = \frac{\beta_\alpha-\beta_0}{\beta_1-\beta_0}$: This is more apparent for

$\beta_0 = 0$ and $\beta_1 = 1$, where this simplifies using $\alpha = \frac{\beta_\alpha - \beta_0}{\beta_1 - \beta_0} = \beta_\alpha$:

$$\begin{aligned}
D_\psi[\beta_0 : \beta_\alpha] &= D_\psi[\beta_1 : \beta_\alpha] \\
&= (1-\beta_\alpha)\psi(0) + \beta_\alpha \psi(1) - \psi(\beta_\alpha) \\
&= (1-\beta_\alpha) \cdot 0 + \beta_\alpha \log p(x) \\
&\quad - \beta_\alpha \log p(x) + (1-\beta_\alpha) D_{\beta_\alpha}[\pi_1(z|x) : \pi_0(z|x)] \\
&= (1-\beta_\alpha) D_{\beta_\alpha}[\pi_1(z|x) : \pi_0(z|x)] \,,
\end{aligned}$$