The generalized Donsker-Varadhan representation

Adam Foster

April 2021

1 Information-theoretic quantities

Throughout machine learning, we have cause to consider the entropy of probability measure p

$$\mathrm{H}\left(p\right) = \mathbb{E}_{p(\mathbf{x})}[-\log p(\mathbf{x})], \tag{1}$$

the KL divergence between two probability measures $p \ll q$

$$\mathrm{KL}\left(p \,\|\, q\right) = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \tag{2}$$

and the mutual information between jointly distributed random variables $\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$

$$I(\mathbf{x}, \mathbf{y}) = \mathrm{KL}\left(p(\mathbf{x}, \mathbf{y}) \,\|\, p(\mathbf{x})p(\mathbf{y})\right). \tag{3}$$

These are foundational quantities in information theory (Shannon, 1948), Bayesian experimental design (Lindley, 1956) and deep learning (Linsker, 1988). A key result in information theory is the following.

Theorem 1 (Gibbs' Inequality)**.** *For any probability measures $p \ll q$,* $\mathrm{KL}\left(p \,\|\, q\right) \geq 0$.

2 The Donsker-Varadhan representation

An important lower bound on the KL divergence is the Donsker-Varadhan (DV) representation.

Theorem 2 (Donsker and Varadhan (1975))**.** *Let $p \ll q$ be probability measures on $\mathcal{X}$, then*

$$\mathrm{KL}\left(p \,\|\, q\right) = \sup_{T:\mathcal{X}\to\mathbb{R} \text{ measurable}} \mathbb{E}_{p(\mathbf{x})}[T(\mathbf{x})] - \log\left(\mathbb{E}_{q(\mathbf{x})}\left[\exp(T(\mathbf{x}))\right]\right) \tag{4}$$

One important bound that can be obtained as a consequence of the Donsker-Varadhan representation is the following.

Corollary 3 (Barber and Agakov (2003))**.** *Let $q(\mathbf{y}|\mathbf{x})$ be a conditional distribution. Then*

$$I(\mathbf{x}, \mathbf{y}) \geq \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{q(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] \tag{5}$$

Proof. Since mutual information is defined as a KL divergence, the DV representation is applicable. Let $T(\mathbf{x}, \mathbf{y}) = \log q(\mathbf{x}, \mathbf{y})/p(\mathbf{y})$ in Theorem 2. We have

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}[q(\mathbf{y}|\mathbf{x})/p(\mathbf{y})] = 1 \tag{6}$$

so the bound is *self-normalized.* The result follows. □

The Barber-Agakov bound can be written as

$$I(\mathbf{x}, \mathbf{y}) \geq \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \left[\log q(\mathbf{y}|\mathbf{x})\right] + \mathrm{H}\left(p(\mathbf{y})\right) \tag{7}$$

which can be helpful in cases in which the $\mathrm{H}\left(p(\mathbf{y})\right)$ term is unknown but also unneeded for e.g. gradient estimation. Another bound, that appears in Nguyen et al. (2010); Nowozin et al. (2016); Belghazi et al. (2018) has a connection to the theory of f-divergences. Applying the inequality $\log x \leq e^{-1}x$ to Theorem 2 gives the **NWJ bound**

$$I(\mathbf{x}, \mathbf{y}) \geq \mathbb{E}_{p(\mathbf{x})}[T(\mathbf{x})] - e^{-1}\mathbb{E}_{q(\mathbf{x})}\left[\exp(T(\mathbf{x}))\right]. \tag{8}$$

An advantage of this looser bound is that it can be directly estimated by samples.

3 A generalization of the Donsker-Varadhan representation

To generalize Theorem 2, suppose we extend the sample space to $\mathcal{X} \times \mathcal{S}$, where $\mathcal{S}$ represents 'side-information'. Suppose we have a conditional distribution $p(\mathbf{s}|\mathbf{x})$. Then we can extend the Donsker-Varadhan representation as follows.

Theorem 4 (Generalized Donsker-Varadhan representation)**.** *Under the assumptions of Theorem 2, let* $p(\mathbf{s}|\mathbf{x})$ *be a valid conditional distribution for each* $\mathbf{x} \in \mathcal{X}$*. Then,*

$$\mathrm{KL}\left(p \,\|\, q\right) = \sup_{U:\mathcal{X}\times\mathcal{S}\to\mathbb{R} \text{ measurable}} \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[U(\mathbf{x}, \mathbf{s})] - \log\left(\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\left[\exp(U(\mathbf{x}, \mathbf{s}))\right]\right) \tag{9}$$

Proof. Since any function $T : \mathcal{X} \to \mathbb{R}$ can be extended to a new function on $\mathcal{X} \times \mathcal{S}$ by ignoring the side information, Theorem 2 immediately tells us that

$$\mathrm{KL}\left(p \,\|\, q\right) \leq \sup_{U:\mathcal{X}\times\mathcal{S}\to\mathbb{R} \text{ measurable}} \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[U(\mathbf{x}, \mathbf{s})] - \log\left(\mathbb{E}_{q(\mathbf{x})p(\mathbf{s},\mathbf{x})}\left[\exp(U(\mathbf{x}, \mathbf{s}))\right]\right). \tag{10}$$

To prove the $\geq$ inequality, we consider some measurable $U : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$. We have

$$\mathrm{KL}\left(p \,\|\, q\right) = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \tag{11}$$

$$= \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\left[\log \frac{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\right] \tag{12}$$

define $V(\mathbf{x}, \mathbf{s}) = \exp(U(\mathbf{x}, \mathbf{s}))/\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\exp(U(\mathbf{x}, \mathbf{s}))]$

$$= \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\left[\log \frac{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})V(\mathbf{x}, \mathbf{s})}\right] + \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\log V(\mathbf{x}, \mathbf{s})] \tag{13}$$

now note that by definition of V, $\int_{\mathcal{X}\times\mathcal{S}} q(\mathbf{x})p(\mathbf{x}|\mathbf{s})V(\mathbf{x}, \mathbf{s}) = 1$, so $q(\mathbf{x})p(\mathbf{x}|\mathbf{s})V(\mathbf{x}, \mathbf{s})$ is a probability measure

$$= \mathrm{KL}\left(p(\mathbf{x})p(\mathbf{s}|\mathbf{x}) \,\|\, q(\mathbf{x})p(\mathbf{s}|\mathbf{x})V(\mathbf{x}|\mathbf{s})\right) + \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\log V(\mathbf{x}, \mathbf{s})] \tag{14}$$

now by Gibbs' Inequality

$$\geq \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\log V(\mathbf{x}, \mathbf{s})] \tag{15}$$

$$= \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[U(\mathbf{x}, \mathbf{s})] - \log\left(\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\left[\exp(U(\mathbf{x}, \mathbf{s}))\right]\right). \tag{16}$$

This completes the proof. □

4 Self-normalized bounds

One particular use of Theorem 4 is for cases in which $\mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\exp(U(\mathbf{x},\mathbf{s}))] = 1$. For such a self-normalized bound, the task of estimating the potentially high-dimensional term $\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\exp(U(\mathbf{x},\mathbf{s}))]$ is removed, and the bound reduces to $\mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[U(\mathbf{x},\mathbf{s})]$ for which unbiased estimators can be constructed directly from samples.

Theorem 5 (Self-normalized KL bound)**.** *Let $k : \mathcal{X} \to \mathbb{R}$ be any measurable function. Then we have the following bound on the KL divergence*

$$\mathrm{KL}\,(p \,\|\, q) \leq \mathbb{E}_{p(\mathbf{x}_1)q(\mathbf{x}_2)\dots q(\mathbf{x}_m)}\left[\log \frac{\exp(k(\mathbf{x}_1))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right]. \tag{17}$$

Proof. We apply Theorem 4 with $\mathbf{x} = \mathbf{x}_1$, $\mathcal{S} = \mathcal{X}^{m-1}$, $\mathbf{s} = (\mathbf{x}_2, ..., \mathbf{x}_m)$ and $p(\mathbf{s}|\mathbf{x}) = q(\mathbf{x}_2) \cdot ... \cdot q(\mathbf{x}_m)$ is independent of $\mathbf{x}_1$. We have

$$U(\mathbf{x},\mathbf{s}) = \log \frac{\exp(k(\mathbf{x}))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))} \tag{18}$$

To apply the theorem, we consider

$$\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\exp(U(\mathbf{x},\mathbf{s}))] = \mathbb{E}_{q(\mathbf{x}_1)\dots q(\mathbf{x}_m)}\left[\frac{\exp(k(\mathbf{x}))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right]. \tag{19}$$

Since the $\mathbf{x}_1, ..., \mathbf{x}_m$ are all equal in distribution, we can replace the index of the sample used in the numerator by any $j \in \{1, ..., m\}$

$$= \mathbb{E}_{q(\mathbf{x}_1)\dots q(\mathbf{x}_m)}\left[\frac{\exp(k(\mathbf{x}_j))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right] \tag{20}$$

we can take the mean over all possible values of j

$$= \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{q(\mathbf{x}_1)\dots q(\mathbf{x}_m)}\left[\frac{\exp(k(\mathbf{x}_j))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right] \tag{21}$$

now by linearity of the expectation we have

$$= \mathbb{E}_{q(\mathbf{x}_1)\dots q(\mathbf{x}_m)}\left[\frac{\frac{1}{m}\sum_{j=1}^{m}\exp(k(\mathbf{x}_j))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right] \tag{22}$$

$$= 1. \tag{23}$$

Thus the bound is self-normalized and the result follows. □

We note that this bound cannot typically recover the KL divergence, because

$$\log \frac{\exp(k(\mathbf{x}_1))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))} \leq \log \frac{\exp(k(\mathbf{x}))}{\frac{1}{m}\exp(k(\mathbf{x}))} = \log m. \tag{24}$$

We can apply a related idea to mutual information. The following theorem provides a self-normalized bound on $I(\mathbf{x},\mathbf{y})$ that is closely related to the popular InfoNCE (van den Oord et al., 2018) bound.

Theorem 6 (Self-normalized information bound)**.** *Let $k : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any measurable function. Then we have the following bound on the mutual information*

$$I(\mathbf{x},\mathbf{y}) \leq \mathbb{E}_{p(\mathbf{x}_1,\mathbf{y}_1)p(\mathbf{x}_2)\dots p(\mathbf{x}_m)}\left[\log \frac{\exp(k(\mathbf{x}_1,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i,\mathbf{y}_1))}\right]. \tag{25}$$

Proof. Since $I(\mathbf{x},\mathbf{y}) = \mathrm{KL}\left(p(\mathbf{x},\mathbf{y}) \,\|\, p(\mathbf{x})p(\mathbf{y})\right)$, we can apply Theorem 4. We set $\mathcal{S} = \mathcal{X}^{m-1}$ and $\mathbf{s} = (\mathbf{x}_2, ..., \mathbf{x}_m)$. We have

$$U((\mathbf{x}_1,\mathbf{y}_1),\mathbf{s}) = \log \frac{\exp(k(\mathbf{x}_1,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i,\mathbf{y}_1))}. \tag{26}$$

To show that this bound is self-normalized, we consider

$$\mathbb{E}_{p(\mathbf{x}_1)p(\mathbf{y}_1)p(\mathbf{s})}[\exp(U((\mathbf{x}_1,\mathbf{y}_1),\mathbf{s}))] = \mathbb{E}_{p(\mathbf{x}_1)\dots p(\mathbf{x}_m)p(\mathbf{y}_1)}\left[\frac{\exp(k(\mathbf{x}_1,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i,\mathbf{y}_1))}\right], \tag{27}$$

for any $\ell \in \{1, ..., m\}$, we have

$$= \mathbb{E}_{p(\mathbf{x}_1)\dots p(\mathbf{x}_m)p(\mathbf{y}_1)}\left[\frac{\exp(k(\mathbf{x}_\ell,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i,\mathbf{y}_1))}\right] \tag{28}$$

since the $\mathbf{x}_i$ are all equal in distribution. Then,

$$= \frac{1}{m}\sum_{\ell=1}^m \mathbb{E}_{p(\mathbf{x}_1)\dots p(\mathbf{x}_m)p(\mathbf{y}_1)}\left[\frac{\exp(k(\mathbf{x}_\ell,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i,\mathbf{y}_1))}\right] \tag{29}$$

$$= \mathbb{E}_{p(\mathbf{x}_1)\dots p(\mathbf{x}_m)p(\mathbf{y}_1)}\left[\frac{\frac{1}{m}\sum_{\ell=1}^m \exp(k(\mathbf{x}_\ell,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i,\mathbf{y}_1))}\right] \tag{30}$$

$$= 1. \tag{31}$$

This completes the proof. ☐

Finally, it is possible to change the distribution that is used to generate $\mathbf{s}$ as long as we compensate with importance weighting. The following theorem gives a bound that is closely connected to the likelihood-free Adaptive Contrastive Estimation bound of Foster et al. (2020) eq. (14).

Theorem 7 (Importance weighted self-normalized information bound)**.** *Let $k : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ be any measurable function. Consider a conditional distribution $q(\mathbf{x}'|\mathbf{y})$ on $\mathcal{X}$. Then we have the following bound on the mutual information*

$$I(\mathbf{x},\mathbf{y}) \le \mathbb{E}_{p(\mathbf{x}_1,\mathbf{y}_1)q(\mathbf{x}_2|\mathbf{y}_1)\dots q(\mathbf{x}_m|\mathbf{y}_1)}\left[\log \frac{\exp(k(\mathbf{x}_1,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i,\mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right]. \tag{32}$$

Proof. Following the same strategy as the previous two proofs, we consider

$$\mathbb{E}_{p(\mathbf{x}_1)p(\mathbf{y}_1)p(\mathbf{s})}[\exp(U((\mathbf{x}_1,\mathbf{y}_1),\mathbf{s}))] = \mathbb{E}_{p(\mathbf{x}_1)p(\mathbf{y}_1)q(\mathbf{x}_{2:m}|\mathbf{y}_1)}\left[\frac{\exp(k(\mathbf{x}_1,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i,\mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right] \tag{33}$$

$$= \mathbb{E}_{p(\mathbf{y}_1)q(\mathbf{x}_{1:m}|\mathbf{y}_1)}\left[\frac{\frac{\exp(k(\mathbf{x}_1,\mathbf{y}_1))p(\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{y}_1)}}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i,\mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right] \tag{34}$$

for any $\ell \in \{1, ..., m\}$, we have

$$= \mathbb{E}_{p(\mathbf{y}_1)q(\mathbf{x}_{1:m}|\mathbf{y}_1)}\left[\frac{\frac{\exp(k(\mathbf{x}_\ell,\mathbf{y}_1))p(\mathbf{x}_\ell)}{q(\mathbf{x}_\ell|\mathbf{y}_1)}}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i,\mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right] \tag{35}$$

since the $\mathbf{x}_i$ are all now equal in distribution. Then,

$$= \frac{1}{m} \sum_{\ell=1}^{m} \mathbb{E}_{p(\mathbf{y}_1)q(\mathbf{x}_{1:m}|\mathbf{y}_1)} \left[\frac{\frac{\exp(k(\mathbf{x}_\ell,\mathbf{y}_1))p(\mathbf{x}_\ell)}{q(\mathbf{x}_\ell|\mathbf{y}_1)}}{\frac{1}{m}\sum_{i=1}^{m} \frac{\exp(k(\mathbf{x}_i,\mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}} \right] \tag{36}$$

$$= \mathbb{E}_{p(\mathbf{y}_1)q(\mathbf{x}_{1:m}|\mathbf{y}_1)} \left[\frac{\frac{1}{m}\sum_{\ell=1}^{m} \frac{\exp(k(\mathbf{x}_\ell,\mathbf{y}_1))p(\mathbf{x}_\ell)}{q(\mathbf{x}_\ell|\mathbf{y}_1)}}{\frac{1}{m}\sum_{i=1} \frac{\exp(k(\mathbf{x}_i,\mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}} \right] \tag{37}$$

$$= 1. \tag{38}$$

This completes the proof. □

A limitation of this bound is that we need to know the density $p(\mathbf{x})$.

References

David Barber and Felix Agakov. The IM algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16:201–208, 2003.

Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R Devon Hjelm, and Aaron Courville. MINE: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. volume 108 of *Proceedings of Machine Learning Research*, pages 2959–2969, Online, 26–28 Aug 2020. PMLR.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11): 5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.