Data Processing Inequality  Function-Space Variational Inference

Andreas Kirsch
University of Oxford[–2023]

Data Processing Inequalities

TL;DR

Informally, the **Data Processing Inequality (DPI)** states that processing data stochastically can only reduce information. Formally, for distributions $q(\Theta)$ and $p(\Theta)$ over a random variable Θ and a stochastic mapping $Y = f(\Theta)$, the Kullback-Leibler DPI states:

$$D_{KL}(q(\Theta) \parallel p(\Theta)) \geq D_{KL}(q(Y) \parallel p(Y)).$$

Equality holds when $D_{KL}(q(\Theta \mid Y) \parallel p(\Theta \mid Y)) = 0$.

The data processing inequality states that if two random variables are transformed in this way, they cannot become easier to tell apart.

"Understanding Variational Inference in Function-Space", Burt et al. (2021)

Example: Image Processing

Consider an image processing pipeline where X is the original image, Y is a compressed version, and Z is Y after adding blur and pixelation. The DPI tells us that the mutual information $I[X;Y] \geq I[X;Z]$, as each processing step results in information loss.

Jenson-Shannon Divergence DPI

The Jensen-Shannon divergence (JSD) makes the KL divergence symmetric. For:

$$f(x) = \frac{p(x) + q(x)}{2}$$

$$D_{JSD}(p(x) \parallel q(x)) = \frac{1}{2} D_{KL}(p(x) \parallel f(x)) + \frac{1}{2} D_{KL}(q(x) \parallel f(x)).$$

The square root of the Jensen-Shannon divergence, the *Jensen-Shannon distance*, is symmetric, satisfies the triangle inequality, and is hence a metric.

For $p(x)$ and $q(x)$ and a shared transition function $f(y \mid x)$ for the model $X \to Y$, we apply the KL DPI twice and obtain the JSD DPI:

$$D_{JSD}(p(X) \parallel q(X)) \geq D_{JSD}(p(Y) \parallel q(Y)).$$

Chain Rule of the KL Divergence

An important property of the KL divergence is the chain rule:

$$\begin{aligned} &D_{KL}(q(Y_n, ...) \parallel p(Y_n, ...)) \\ &\quad = D_{KL}(q(Y_n \mid Y_{n-1}, ...) \parallel p(Y_n \mid Y_{n-1}, ...)) \\ &\qquad + D_{KL}(q(Y_{n-1}, ...) \parallel p(Y_{n-1}, ...)) \\ &= ... = \sum_{i=1}^{n} D_{KL}(q(Y_i \mid Y_{i-1}, ...) \parallel p(Y_i \mid Y_{i-1}, ...)). \end{aligned}$$

This also yields a **chain inequality**:

$$\begin{aligned} D_{KL}(q(Y_n, ...) \parallel p(Y_n, ...)) &\geq D_{KL}(q(Y_{n-1}, ...) \parallel p(Y_{n-1}, ...)) \\ &\geq ... \geq D_{KL}(q(Y_1) \parallel p(Y_1)), \end{aligned}$$

by repeatedly applying the chain rule and dropping the conditional term.

Mutual Information DPI

For any Markov chain $Z \to X \to Y$ with $f(z, x, y) = f(z)f(x \mid z)f(y \mid x)$ for any distribution $f(z)$, we have:

$$\begin{aligned} I[X;Z] &= D_{KL}(f(X \mid Z) \parallel f(X)) \\ &= \mathbb{E}_{f(z)} \left[D_{KL}(f(X \mid z) \parallel f(X)) \right] \\ &\overset{(1)}{\geq} \mathbb{E}_{f(z)} \left[D_{KL}(f(Y \mid z) \parallel f(Y)) \right] \\ &= D_{KL}(f(Y \mid Z) \parallel f(Y)) \\ &= I[Y;Z], \end{aligned}$$

where (1) follows from the KL DPI.

Proof of the DPI

Using the chain rule of the KL divergence twice:

$$\begin{aligned} &D_{KL}(p(X) \parallel q(X)) + \underbrace{D_{KL}(p(Y \mid X) \parallel q(Y \mid X))}_{=D_{KL}(f(Y \mid X) \parallel f(Y \mid X))=0} \\ &= D_{KL}(p(X,Y) \parallel q(X,Y)) \\ &= D_{KL}(p(Y) \parallel q(Y)) + \underbrace{D_{KL}(p(X \mid Y) \parallel q(X \mid Y))}_{\geq 0} \\ &\geq D_{KL}(p(Y) \parallel q(Y)). \end{aligned}$$

We have equality exactly when $p(x \mid y) = q(x \mid y)$ for (almost) all x, y.

Function-Space Variational Inference

TL;DR

Function-space variational inference (FSVI) is a principled approach to Bayesian inference that respects the inherent symmetries and equivalences in overparameterized models. It focuses on approximating the meaningful posterior $p([\boldsymbol{\theta}] \mid \mathcal{D})$ over prediction equivalence classes of the parameters while avoiding the complexities of explicitly constructing and working with equivalence classes. The FSVI-ELBO regularizes towards a data prior using the KL DPI:

$$\mathbb{E}_{q(\boldsymbol{\theta})} \left[-\log p(\mathcal{D} \mid \boldsymbol{\theta}) \right] + D_{KL}(q(Y... \mid \boldsymbol{x}...) \parallel p(Y... \mid \boldsymbol{x}...)),$$

(unlike in regular variational inference, where we regularize towards a parameter prior $D_{KL}(q(\Theta) \parallel p(\Theta))$).

(Regular) Variational Inference & ELBO

The Bayesian posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ is approximated with a variational distribution $q(\boldsymbol{\theta})$ by minimizing $D_{KL}(q(\Theta) \parallel p(\Theta \mid \mathcal{D}))$. Dropping constant (intractable) terms yields a simplified tractable objective, which **upper** bounds the information content $-\log p(\mathcal{D})$ of the data $\mathcal{D}$:

$$\begin{aligned} &0 \leq D_{KL}(q(\Theta) \parallel p(\Theta \mid \mathcal{D})) \\ &= D_{KL}(q(\Theta) \parallel \frac{p(\mathcal{D} \mid \Theta)\, p(\Theta)}{p(\mathcal{D})}) \\ &= \underbrace{\mathbb{E}_q \left[-\log p(\mathcal{D} \mid \Theta) \right] + D_{KL}(q(\Theta) \parallel p(\Theta))}_{\text{Evidence Bound (Simplified Objective)}} \\ &\quad - (\underbrace{-\log p(\mathcal{D})}_{\text{(neg. log) Evidence}}). \end{aligned}$$

This is equivalent to the **evidence lower bound (ELBO)**.

Parameter Symmetries

Deep neural networks have many parameter symmetries: for example, in a convolutional neural network, we could swap channels without changing the predictions. $\implies$ *We are not interested in these symmetries but only differing predictions.*

Equivalence Classes

Equivalence classes group together parameters that lead to the same predictions on a (test) set of data:

$$[\boldsymbol{\theta}] \triangleq \{\boldsymbol{\theta}' : f(x; \boldsymbol{\theta}) = f(x; \boldsymbol{\theta}) \quad \forall x\}.$$

Crucially, *different domains for $\boldsymbol{x}$ will induce different equivalence classes.*

Consistency of Equivalence Classes with Bayesian Inference

Any distribution over the parameters $p(\boldsymbol{\theta})$ induces a distribution $p([\boldsymbol{\theta}])$ over the equivalence classes, which is consistent with Bayesian inference:

$$p([\boldsymbol{\theta}]) \triangleq \sum_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}]} p(\boldsymbol{\theta}'),$$

that is, $[\boldsymbol{\theta}]$ commutes with Bayesian inference:

$$p([\boldsymbol{\theta}] \mid \mathcal{D}) = \sum_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}]} p(\boldsymbol{\theta}' \mid \mathcal{D}) \Leftrightarrow [\Theta \mid \mathcal{D}] = [\Theta] \mid \mathcal{D}.$$

This commutative property is a general characteristic of applying (stochastic) functions to random variables.

$$\begin{array}{ccc} \Theta & \xrightarrow{\cdot \mid \mathcal{D}} & \Theta \mid \mathcal{D} \\ \downarrow{\scriptstyle [\cdot]} & & \downarrow{\scriptstyle [\cdot]} \\ [\Theta] & \xrightarrow{\cdot \mid \mathcal{D}} & [\Theta] \mid \mathcal{D} \end{array}$$

Commutative diagram for the equivalence classes.

Equality in the Infinite Data Limit

$$\begin{aligned} D_{KL}(q(\Theta) \parallel p(\Theta)) &\geq D_{KL}(q([\Theta]) \parallel p([\Theta])) \\ &\geq D_{KL}(q(Y_n... \mid \boldsymbol{x}_n...) \parallel p(Y_n... \mid \boldsymbol{x}_n...)). \end{aligned}$$

Unless there are no parameter symmetries, the **first inequality will not be tight** ($D_{KL}(q(\Theta \mid [\Theta]) \parallel p(\Theta \mid [\Theta])) > 0$). The **second inequality will be tight**: to see this, we need $D_{KL}(q([\Theta] \mid Y_n, \boldsymbol{x}_n, ...) \parallel p([\Theta] \mid Y_n, \boldsymbol{x}_n, ...)) \to 0$ for $n \to \infty$. The term *converges* as it is monotonically increasing and bounded by $D_{KL}(q([\Theta]) \parallel p([\Theta]))$ from above. Further, thanks to Bernstein von Mises' theorem, we have $\to 0$ and thus:

$$\begin{aligned} &D_{KL}(q([\Theta]) \parallel p([\Theta])) = \\ &\qquad = \sup_{n \in \mathbb{N}} D_{KL}(q(Y_n, ... \mid \boldsymbol{x}_n, ...) \parallel p(Y_n, ... \mid \boldsymbol{x}_n, ...)). \end{aligned}$$

Bernstein von Mises' Theorem

BvM states that a posterior distribution converges to the maximum likelihood estimate (MLE) as the number of data points tends to infinity *as long as the model parameters are identifiable, that is the true parameters we want to learn are unique, and they have support.* This is true for $[\Theta]$ as $[\Theta]$ is a unique parameterization of the same distribution as Θ, and we can ensure support.

Function-Space Variational Inference & ELBO

FSVI's ELBO is just the regular ELBO but for $[\Theta]$ and uses further approximations via chain inequalities:

$$\begin{aligned} H[\mathcal{D}] &\leq H[\mathcal{D}] + D_{KL}(q([\Theta]) \parallel p([\Theta] \mid \mathcal{D})) \\ &= H[\mathcal{D}] + D_{KL}(q([\Theta]) \parallel \frac{p(\mathcal{D} \mid [\Theta])\, p([\Theta])}{p(\mathcal{D})}) \\ &= \mathbb{E}_{q([\boldsymbol{\theta}])} \left[-\log p(\mathcal{D} \mid [\boldsymbol{\theta}]) \right] + D_{KL}(q([\Theta]) \parallel p([\Theta])). \end{aligned}$$

Then, we can apply the chain rule together with BvM:

$$\begin{aligned} &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[-\log p(\mathcal{D} \mid \boldsymbol{\theta}) \right] + \sup_n D_{KL}(q(\cdot) \parallel p(\cdot))[Y_n... \mid \boldsymbol{x}_n...] \\ &\geq \mathbb{E}_{q(\boldsymbol{\theta})} \left[-\log p(\mathcal{D} \mid \boldsymbol{\theta}) \right] + D_{KL}(q(\cdot) \parallel p(\cdot))[Y_n... \mid \boldsymbol{x}_n...] \quad \forall n. \end{aligned}$$

Full Blog Post

Burt, D. R., Ober, S. W., Garriga-Alonso, A., and van der Wilk, M. (2021). Understanding variational inference in function-space. In *Third Symposium on Advances in Approximate Bayesian Inference*.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Rudner, T. G. J., Chen, Z., Teh, Y. W., and Gal, Y. (2022). Tractable function-space variational inference in bayesian neural networks. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.