Bridging the Data Processing Inequality and 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 $\mathrm{q}(\boldsymbol{\Theta})$ and $\mathrm{p}(\boldsymbol{\Theta})$ over a random variable $\boldsymbol{\Theta}$ and a stochastic mapping $Y = \mathrm{f}(\boldsymbol{\Theta})$, the DPI is expressed as:

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

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

> *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 $\mathrm{I}[X;Y] \geq \mathrm{I}[X;Z]$, as each processing step results in information loss.

Jenson-Shannon Divergence DPI

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

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

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

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

For $\mathrm{p}(x)$ and $\mathrm{q}(x)$ and shared transition function $\mathrm{f}(y \mid x)$ for the model $X \to Y$:

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

Mutual Information DPI

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

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

where (1) follows from the KL DPI.

Chain Rule of the Divergence

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

$$\mathrm{D_{KL}}(\mathrm{q}(Y_n, \ldots) \parallel \mathrm{p}(Y_n, \ldots)) = \sum_{i=1}^{n} \mathrm{D_{KL}}(\mathrm{q}(Y_i \mid Y_{i-1}, \ldots) \parallel \mathrm{p}(Y_i \mid Y_{i-1}, \ldots)).$$

This chain rule also yields a **chain inequality**:

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

where we start from the KL DPI and then use the chain rule.

Proof of the DPI

Using the chain rule of the KL divergence twice:

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

We have equality exactly when $\mathrm{p}(x \mid y) = \mathrm{q}(x \mid y)$ 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 $\mathrm{p}([\boldsymbol{\theta}] \mid \mathcal{D})$ while avoiding the complexities of explicitly constructing and working with equivalence classes. The FSVI-ELBO regularizes towards a data prior:

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

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

(Regular) Variational Inference & ELBO

We approximate the Bayesian posterior $\mathrm{p}(\boldsymbol{\theta} \mid \mathcal{D})$ with a variational distribution $\mathrm{q}(\boldsymbol{\theta})$ by minimizing $\mathrm{D_{KL}}(\mathrm{q}(\boldsymbol{\Theta}) \parallel \mathrm{p}(\boldsymbol{\Theta} \mid \mathcal{D}))$ and dropping constant (intractable) terms to obtain a simplified objective, which also yields an information-theoretic **upper** bound on the information content $-\log \mathrm{p}(\mathcal{D})$ of the data $\mathcal{D}$:

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

The negative of this bound is called the **evidence lower bound (ELBO)**.

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

Equality in the Infinite Data Limit

Using the DPI:

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

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

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

Parameter Symmetries

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

Equivalence Classes

We can use **equivalence classes** to group together parameters that lead to the same predictions on a (test) set of data:

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

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

Consistency of Equivalence Classes with Bayesian Inference

Any distribution over the parameters $\mathrm{p}(\boldsymbol{\theta})$ induces a distribution $\hat{\mathrm{p}}([\boldsymbol{\theta}])$ over the equivalence classes:

$$\hat{\mathrm{p}}([\boldsymbol{\theta}]) \triangleq \sum_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}]} \mathrm{p}(\boldsymbol{\theta}').$$

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

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

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

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 that they have support*, which is true for $[\boldsymbol{\Theta}]$.

Function-Space Variational Inference & ELBO

FSVI's ELBO is just the regular ELBO but for $[\boldsymbol{\Theta}]$ and approximations via chain rule of the DPI:

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

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

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

More Info

More References

[1] Thomas M Cover. *Elements of information theory.* John Wiley & Sons, 1999.

[2] Tim G. J. Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in bayesian neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.