Machine Learning using the Variational Predictive Information Bottleneck with a Validation Set

Sayandev Mukherjee
CableLabs, Sunnyvale, CA
s.mukherjee@cablelabs.com
November 24, 2021

1 Abstract

Zellner [1] modeled statistical inference in terms of information processing and postulated the Information Conservation Principle (ICP) between the input and output of the information processing block, showing that this yielded Bayesian inference as the optimum information processing rule. Recently, Alemi [2] reviewed Zellner’s work in the context of machine learning and showed that the ICP could be seen as a special case of a more general optimum information processing criterion, namely the predictive information bottleneck objective. However, [2] modeled machine learning as using training and test data sets only, and did not account for the use of a validation data set during training. The present note is an attempt to extend Alemi’s information processing formulation of machine learning, and the predictive information bottleneck objective for model training, to the widely-used scenario where training utilizes not only a training but also a validation data set.

2 Review of Information Processing formulation of Machine Learning

2.1 Introduction and Notation

We will use Alemi’s formulation and notation from [2], with some additional detail for clarity. Consider a data generating process $\phi$ with distribution (PMF or PDF, depending on whether $\phi$ is discrete or continuous-valued, respectively) $p(\phi)$, which generates the features $x$ according to the distribution $p(x|\phi)$. We collect $N$ samples of $x$ in the training set $x_P = \{x_1, x_2, \ldots, x_N\}$, with
the choice of subscript ‘P’ emphasizing that these are past observations. Depending on whether we are testing the performance of a trained model on a test set or deploying a trained model in a production environment to perform inference, we may have a finite or (potentially) infinite set \(x_F = \{x_{N+1}, \ldots\}\) of future (i.e., not seen during training) samples of \(x\) from the same process (also emphasized by the choice of subscript ‘F’).

### 2.2 The Predictive Information Bottleneck Objective for Model Training

Model training or “learning” is the extraction of the model parameters \(\theta\) from the training set \(x_P\). Viewed from the perspective of information processing, we may see model training as computing, and sampling from, the distribution \(p(\theta | x_P)\).

Again from the perspective of information processing, the trained model may be evaluated by how much information the training-derived representation \(p(\theta | x_P)\) captures about future samples \(x_F\). In other words, we want to find the \(p(\theta | x_P)\) that maximizes the mutual information \(I(\theta; x_F)\):

\[
\max_{p(\theta | x_P)} I(\theta; x_F),
\]

where for any two random variables \(x\) and \(y\), their mutual information \(I(x; y)\) is defined as the Kullback-Leibler distance between their joint distribution \(p(x, y)\) and the product of their marginal distributions \(p(x)\) and \(p(y)\):

\[
I(x; y) = D_{KL}(p(x, y) \parallel p(x)p(y)) = \mathbb{E}_{p(x,y)} \left[ \log \frac{p(x,y)}{p(x)p(y)} \right] = I(y; x).
\]

Alemi \[2\] proposes to limit the complexity of the model representation obtained from training by incorporating a bottleneck requirement on the mutual information \(I(\theta; x_P)\). This yields the following predictive information bottleneck objective on training:

\[
\max_{p(\theta | x_P)} I(\theta; x_F) \quad \text{subject to} \quad I(\theta; x_P) = I_0. \tag{1}
\]

Applying a Lagrange multiplier \(\lambda\) to the constraint on \(I(\theta; x_P)\), \(1\) can be rewritten as the unconstrained optimization problem

\[
\max_{\lambda, p(\theta | x_P)} [I(\theta; x_F) - \lambda I(\theta; x_P)]. \tag{2}
\]

\[1\]Note that we assume that \(I(\theta; x_P)\) is finite and not a constant. For this, we require that \(\theta\) not be a deterministic function of \(x_P\). This requirement is satisfied for a deep learning model trained by gradient descent (because of random initialization) or stochastic gradient descent (because of both random initialization and randomized selection of minibatch entries by replacement from \(x_P\)).
Note that there is no constraint on the sign of $\lambda$. Next, from the Markov property of the information processing chain $x_F \leftarrow \phi \rightarrow x_P \rightarrow \theta$, we have $I(\theta; x_F, x_P) = I(\theta; x_P)$, so we have

\[
I(\theta; x_F) - I(\theta; x_P) = I(\theta; x_F, x_P) - I(\theta; x_P | x_F) - I(\theta; x_P)
\]

\[
\quad = -I(\theta; x_P | x_F), \tag{3}
\]

where in the first step we have used the identity

\[
I(x; y, z) = I(x; y) + I(x; z | y). \tag{4}
\]

Combining (2) and (3) then yields the following unconstrained optimization problem equivalent to the predictive information bottleneck objective (1):

\[
\min_{\beta, p(\theta | x_P)} \left[ I(\theta; x_P | x_F) - \beta I(\theta; x_P) \right], \tag{5}
\]

where $\beta = 1 - \lambda$. Since (5) cannot be solved directly, Alemi employs two variational approximations that are described next.

### 2.3 Variational Approximations to the Predictive Information Bottleneck Objective

- Treat $\theta$ as unobserved variables and use a variational approximation $q(\theta)$ to the true distribution $p(\theta | x_F)$, where $q(\cdot)$ is a distribution chosen independent of $x_F$. Denoting by $\mathbb{E}$ the expectation with respect to (the true distributions of) all the random variables $\theta, x_P, \phi, x_F$, we have (recalling that $\theta$ and $x_F$ are conditionally independent given $x_P$):

\[
I(\theta; x_P | x_F) = \mathbb{E} \left[ \log \frac{p(\theta, x_P | x_F)}{p(\theta | x_F) p(x_P | x_F)} \right]
\]

\[
\quad = \mathbb{E} \left[ \log \frac{p(\theta | x_P, x_F)}{p(\theta | x_F)} \right] = \mathbb{E} \left[ \log \frac{p(\theta | x_P)}{p(\theta | x_F)} \right]
\]

\[
\quad = \mathbb{E} \left[ \log \frac{p(\theta | x_P)}{q(\theta)} \right] - \mathbb{E} \left[ \log \frac{p(\theta | x_F)}{q(\theta)} \right] + \mathbb{E} \left[ \log \frac{p(\theta | x_F)}{q(\theta)} \right]
\]

\[
\quad = \mathbb{E} \left[ \log \frac{p(\theta | x_P)}{q(\theta)} \right] - \mathbb{E} \left[ \log \frac{p(\theta | x_F)}{q(\theta)} \right] + \mathbb{E} \left[ D_{KL} \left( p(\cdot | x_F) \parallel q(\cdot) \right) \right]
\]

\[
\quad \leq \mathbb{E} \left[ \log \frac{p(\theta | x_P)}{q(\theta)} \right]. \tag{6}
\]
• Treat the distribution \( p(x_P | \theta) \) as a “likelihood” function and use a variational approximation for it given by \( q(x_P | \theta) \), for example the factorized form \( q(x_P | \theta) = \prod_{x \in x_P} q(x | \theta) \) for some selected distribution \( q(\cdot | \theta) \). Then we can write

\[
I(\theta; x_P) = \mathbb{E} \left[ \log \frac{p(\theta, x_P)}{p(\theta) p(x_P)} \right] = \mathbb{E} \left[ \log \frac{p(x_P | \theta)}{q(x_P | \theta)} \right] - \mathbb{E} \left[ \log p(x_P) + \log q(x_P | \theta) \right] \\
= \mathbb{E} \left[ D_{KL} \left( p(\cdot | \theta) \parallel q(\cdot | \theta) \right) \right] + \mathbb{E} \left[ - \log p(x_P) \right] + \mathbb{E} \left[ \log q(x_P | \theta) \right] \\
\geq 0 + H(x_P) + \mathbb{E} \left[ \log q(x_P | \theta) \right],
\]

where for any random variable \( x \), \( H(x) \) is its entropy:

\[
H(x) = - \sum_x p(x) \log p(x) \quad \text{or} \quad H(x) = - \int p(x) \log p(x) \, dx.
\]

### 2.4 Variational Formulation of Predictive Information Bottleneck Objective

From (6) and (7) we therefore have, for every choice of \( \beta \geq 0 \), variational approximate marginal distribution \( q(\theta) \), and variational approximate likelihood function \( q(x | \theta) \), the following upper bound on the objective function of (5):

\[
I(\theta; x_P | x_F) - \beta I(\theta; x_P) \leq \mathbb{E} \left[ \log \frac{p(\theta | x_P)}{q(\theta)} \right] - \beta \mathbb{E} \left[ \log q(x_P | \theta) \right] - \beta H(x_P).
\]

Note that \( H(x_P) \) is a constant outside our control. Since the exact problem (5) cannot be solved directly, we can simply select \( \beta \geq 0, q(\theta), q(x | \theta) \) based on some external criteria and solve the following problem, whose optimum value yields an upper bound on the exact objective in (5):

\[
\min_{p(\theta | x_P)} \mathbb{E}_{p(\phi)p(x_P | \phi)p(\theta | x_P)} \left[ \log \frac{p(\theta | x_P)}{q(\theta)} \right] - \beta \mathbb{E} \left[ \log q(x_P | \theta) \right] \\
= \min_{p(\theta | x_P)} \mathbb{E}_{p(\phi)p(x_P | \phi)} \left\{ \mathbb{E}_{p(\theta | x_P)} \left[ \log \frac{p(\theta | x_P)}{q(\theta) q(x_P | \theta)^\beta} \right] \right\} \\
= \min_{p(\theta | x_P)} \mathbb{E}_{p(\phi)p(x_P | \phi)} \left[ D_{KL} \left( p(\cdot | x_P) \parallel q(\cdot | x_P) \right) \frac{q(\cdot) q(x_P | \cdot)^\beta}{Z_\beta(x_P)} \right],
\]

where

\[
Z_\beta(x_P) = \sum_\psi q(\psi) q(x_P | \psi)^\beta \quad \text{or} \quad Z_\beta(x_P) = \int q(\psi) q(x_P | \psi)^\beta \, d\psi
\]
and we emphasize that the expectation in (9) is with respect to the true distributions of $\theta, x_P, \phi$. Note that the use of the variational approximation $q(\theta)$ has eliminated the dependence on the distribution of $x_F$.

Finally, it follows from (10) that the optimum distribution $p(\theta | x_P)$ is given by

$$p(\theta | x_P) = \frac{q(\theta) q(x_P | \theta)^\beta}{Z_\beta(x_P)}.$$

For $\beta = 1$, the objective in (10) can be identified as the ICP postulated by Zellner [1], and the optimum $p(\theta | x_P)$ is the Bayesian inference derived from the variational marginal and likelihood $q(\theta)$ and $q(x | \theta)$ respectively:

$$p(\theta | x_P) \propto q(\theta) q(x_P | \theta).$$

### 3 The Modified Predictive Information Bottleneck Objective

#### 3.1 Shortcomings of the Predictive Information Bottleneck Objective (1)

Alemi’s formulation [2] of the predictive information bottleneck objective in (1) has the following shortcomings:

- Recall that in (1), the Lagrange multiplier $\lambda$ may be positive or negative-valued. However, the variational upper bound in (8) does not hold unless $\beta \geq 0$, i.e., $\lambda \leq 1$, though there is no explanation offered in [2] of why only values of $\lambda$ in this range make sense for this problem.

- The quantity $I_0$ is never used after the equality constraint $I(\theta; x_P) = I_0$ is incorporated into the problem formulation in (1). Further, $\beta$ is simply selected as per other criteria, and not set so as to achieve this equality $I(\theta; x_P) = I_0$.

#### 3.2 Predictive Information Bottleneck Objective with Inequality Constraint

The fundamental reason for the above shortcomings is the requirement of equality in the bottleneck constraint.

We therefore propose to modify the predictive information bottleneck objective from (1) to one that has an inequality constraint between $I(\theta; x_P)$, the mutual information of the trained model and the training set, and a pre-selected threshold $I_0$:

$$\max_{p(\theta | x_P)} I(\theta; x_P) \quad \text{subject to} \quad I(\theta; x_P) \geq I_0 \iff I_0 - I(\theta; x_P) \leq 0. \quad (11)$$
Note that this formulation imposes a bottleneck on model performance, such that only models that extract a certain threshold level of information from the training set clear the bottleneck. From the Karush-Kuhn-Tucker (KKT) theorem, the optimization problem (1) now changes to

\[
\max_{\mu \geq 0, p(\theta | x_P)} \left\{ I(\theta; x_F) - \mu \left[ I_0 - I(\theta; x_F) \right] \right\} = \max_{\mu \geq 0, p(\theta | x_P)} \left[ I(\theta; x_F) + \mu I(\theta; x_P) \right] \tag{12}
\]

\[
= \min_{\beta \geq 1, p(\theta | x_P)} \left[ I(\theta; x_F | x_P) - \beta I(\theta; x_P) \right], \tag{13}
\]

where we have used (3) to go from (12) to (13), and \( \beta = 1 + \mu \geq 1 \) because \( \mu \geq 0 \) by definition. Note that this immediately resolves the interpretability issues that arose with the equality bottleneck constraint formulation above. Note also that the objective function of (13) is the same as the left hand side of (8). Then the variational approximations discussed in Sec. 2.3 can be applied as before, resulting in the same variational objective function (10).

Further, the KKT theorem requires that the optimum \( \mu, p(\theta | x_P) \) for (12) satisfy the complementary slackness condition

\[
\mu \left[ I_0 - I(\theta; x_P) \right] = 0, \tag{14}
\]

which implies that if the training step achieves \( I(\theta; x_P) > I_0 \), then the optimum \( \mu \) is zero, i.e., the optimum \( \beta \) in (13) is unity. From the discussion at the end of Sec. 2.4 we have the following result.

**Theorem 1**  **If training step is successful, i.e., we achieve** \( I(\theta; x_P) > I_0 \), **then the optimum information processing for (9) is Bayesian inference:**

\[
p(\theta | x_P) \propto q(\theta) q(x_P | \theta), \tag{15}
\]

where \( q(\theta) \) and \( q(x | \theta) \) are as defined in Sec. 2.3.

**Remark 1**  In a practical scenario where we train, say, a chosen deep learning architecture with stochastic gradient descent over mini-batches sampled from the training set, “success” in training is defined as some measure of performance on the training set, such as average training loss (if the model is being trained for a regression or function approximation application) or classification error (if the model is being trained for classification), exceeding some threshold. In the present information-theoretic framework, this performance threshold will need to be translated into some threshold \( I_0 \) on the mutual information \( I(\theta; x_P) \). Now, \( I(\theta; x_P) \) cannot be computed or even estimated directly, but we can use the variational lower bound (7) on \( I(\theta; x_P) \) as follows:

- Estimate the entropy \( H(x_P) \) of the training set using a source-coding or compression scheme;
• Estimate $L \equiv \mathbb{E}[\log q(x_P \mid \theta)]$ from an average over mini-batches sampled from $x_P$;
• Declare “successful training” when $\hat{H}(x_P) + \hat{L} > I_0$, where the hats denote the estimates.

4 Extension to Model Training with a Validation Set

4.1 Introduction and Notation

The machine learning model is trained on a training set $x_P$. For training classical machine learning models like support vector machines, decision trees, and random forests, which are not as computationally expensive to train as deep learning models, this training set $x_P$ can be used for cross-validation to validate the performance of the model during training in order to guard against overfitting the training set. However, $k$-fold cross-validation requires the model to be trained from scratch $k$ times, which often imposes an unacceptable computational burden when the model is a deep learning one. Thus, when training deep learning models, it is widespread practice to set aside a portion of the available training data as a validation set, use the remainder of the training data as the training set to train the model, and, as the model trains, validate its performance on the validation set.

In the notation of the previous section, if the available training data comprises the $N$ samples $x_1, x_2, \ldots, x_N$, the training set is now $x_P = \{x_1, x_2, \ldots, x_M\}$ where $M < N$, with the rest of the training data comprising the validation set $x_V = \{x_{M+1}, \ldots, x_N\}$.

4.2 Predictive Information Bottleneck Objective with Validation Set

As in Sec. 3.2, we will require the trained model to clear the bottleneck of a threshold on mutual information between the model and the training set:

$$I(\theta ; x_P) \geq I'_0. \quad (16)$$

In addition, we will require the model, after being trained on the training set $x_P$, to clear another threshold of mutual information between the model and the validation set (analogous to the performance requirement on mutual information between the model and the test set $x_F$):

$$I(\theta ; x_V \mid x_P) \geq I''_0. \quad (17)$$

From (4) we see that (16) and (17) together yield

$$I(\theta ; x_P, x_V) \geq I_0 \equiv I'_0 + I''_0, \quad (18)$$
which is equivalent to requiring the bottleneck condition with threshold $I_0 = I'_0 + I''_0$ on the mutual information between the model and the augmented training set $\{x_1, x_2, \ldots, x_N\}$ created from $x_P$ and $x_V$. However, instead of requiring the single condition (18), we shall impose the stricter requirement to satisfy the two conditions (16) and (17).

The new optimization problem is now

$$\max_{p(\theta \mid x_P, x_V)} I(\theta; x_F) \quad \text{subject to} \quad I(\theta; x_P) \geq I'_0 \quad \text{and} \quad I(\theta; x_V \mid x_P) \geq I''_0. \quad (19)$$

In (19) the distribution $p(\theta \mid x_P, x_V)$ incorporates the dependence of the trained model on the validation set. This dependence arises because we train by running a learning algorithm on (a subset of) the training set, validate the (partially) trained model on the validation set, run the learning algorithm again on (another subset of) the training set, validate the (partially) trained model again on the validation set, and so on until the performance of the model on both the training and validation sets satisfies some criteria which we represent here by (16) and (17) respectively.

### 4.3 Modified Predictive Information Bottleneck Objective with Inequality Constraint and Validation Set

From the Markov property of the information processing chain $x_F \leftarrow \phi \rightarrow (x_P, x_V) \rightarrow \theta$, we have $I(\theta; x_F, (x_P, x_V)) = I(\theta; x_P, x_V)$, so we have

$$I(\theta; x_F) = I(\theta; x_F) - I(\theta; x_P) - I(\theta; x_V \mid x_P)$$

$$= I(\theta; x_F) - I(\theta; x_P) - I(\theta; x_P, x_V)$$

$$= I(\theta; x_F, (x_P, x_V)) - I(\theta; x_P, x_V \mid x_F) - I(\theta; x_P, x_V) \quad \text{[from (4)]}$$

$$= -I(\theta; x_P, x_V \mid x_F). \quad (20)$$

As with (11), the KKT theorem lets us rewrite (19) as follows:

$$\max_{\mu \geq 0, \nu \geq 0, p(\theta \mid x_P, x_V)} \left\{ I(\theta; x_F) - \mu \left[ I'_0 - I(\theta; x_P) \right] - \nu \left[ I''_0 - I(\theta; x_V \mid x_P) \right] \right\}$$

$$= \max_{\mu \geq 0, \nu \geq 0, p(\theta \mid x_P, x_V)} \left[ I(\theta; x_F) + \mu I(\theta; x_P) + \nu I(\theta; x_V \mid x_P) \right] \quad (21)$$

$$= \min_{\beta \geq 1, \gamma \geq 1, p(\theta \mid x_P, x_V)} \left[ I(\theta; x_P, x_V \mid x_F) - \beta I(\theta; x_P) - \gamma I(\theta; x_V \mid x_P) \right], \quad (22)$$

where we use (20) to go from (21) to (22), and $\beta = 1 + \mu \geq 1$, $\gamma = 1 + \nu \geq 1$. Further, the
optimum \( \beta, \gamma, p(\theta \mid x_P, x_V) \) must satisfy the complementary slackness conditions

\[
\mu \left[ I_0^\prime - I(\theta; x_P) \right] = 0, \quad (23)
\]

\[
\nu \left[ I_0'' - I(\theta; x_V \mid x_P) \right] = 0. \quad (24)
\]

### 4.4 Variational Approximations

- Treat \( \theta \) as unobserved variables and use a variational approximation \( q(\theta) \) to the true distribution \( p(\theta \mid x_F) \), where \( q(\cdot) \) is a distribution chosen independent of \( x_F \). Following the notation of Sec. 2.3 and the same steps as in the derivation of (6), we obtain

\[
I(\theta; x_P, x_V \mid x_F) \leq \mathbb{E} \left[ \log \frac{p(\theta \mid x_P, x_V)}{q(\theta)} \right]. \quad (25)
\]

- Treat the distribution \( p(x_P \mid \theta) \) as a “likelihood” function and use a variational approximation for it given by \( q(x_P \mid \theta) \), for example the factorized form \( q(x_P \mid \theta) = \prod_{x \in x_P} q(x \mid x_P, \theta) \) for some selected distribution \( q(\cdot \mid x_P, \theta) \). Then we obtain (7):

\[
I(\theta; x_P) \geq H(x_P) + \mathbb{E} [\log q(x_P \mid \theta)].
\]

- Treat the conditional distribution \( p(x_V \mid x_P, \theta) \) as a conditional likelihood function and use a variational approximation for it given by \( q(x_V \mid x_P, \theta) \), for example the factorized form \( q(x_V \mid x_P, \theta) = \prod_{x \in x_V} q(x \mid x_P, \theta) \) for some selected distribution \( q(\cdot \mid x_P, \theta) \). Then we obtain

\[
I(\theta; x_V \mid x_P) \geq 0 + H(x_V \mid x_P) + \mathbb{E} [\log q(x_V \mid x_P, \theta)], \quad (26)
\]
where for any two random variables $x$ and $y$,

$$ H(x \mid y) = -\sum_{x, y} p(x, y) \log p(x \mid y) \quad \text{or} \quad H(x \mid y) = -\int p(x, y) \log p(x \mid y) \, dx \, dy. $$

Note that the process of training on $x_P$ along with periodic validation using the validation set $x_V$ induces a dependence between $x_V$ and $x_P$ through the model $\theta$, hence it is not true in general that $p(x_V \mid x_P, \theta) = p(x_V \mid \theta)$ even if $x_V$ and $x_P$ are assumed independent.

### 4.5 Variational Formulation of the Predictive Information Bottleneck

**Objective with Inequality Constraint and a Validation Set**

From (25), (7), and (26), the objective function in (22) is upper-bounded for each choice of $\beta \geq 1, \gamma \geq 1, q(\theta), q(x \mid \theta), q(x \mid x_P, \theta)$ as follows:

$$ I(\theta; x_P, x_V \mid x_F) - \beta I(\theta; x_P) - \gamma I(\theta; x_V \mid x_P) $$

$$ \leq \mathbb{E} \left[ \log \frac{p(\theta \mid x_P, x_V)}{q(\theta)} - \beta \log q(x_P \mid \theta) - \gamma \log q(x_V \mid x_P, \theta) \right] - \beta H(x_P) - \gamma H(x_V \mid x_P) $$

$$ = \mathbb{E} \left[ \log \frac{p(\theta \mid x_P, x_V)}{q(\theta)} - \beta \log q(x_P \mid \theta) - \gamma \log q(x_V \mid x_P, \theta) \right] $$

$$ - \beta H(x_P) - \gamma H(x_V) + \gamma I(x_P; x_V), $$

where in the final step we use the identity that for any two random variables $x$ and $y$,

$$ I(x; y) = H(x) - H(x \mid y). $$

We treat the last three terms on the right hand side of (27) as constants outside our control. Since the exact problem (22) cannot be solved directly, we simply select $\beta \geq 1, \gamma \geq 1, q(\theta), q(x \mid \theta), q(x \mid x_P, \theta)$ based on some external criteria and solve the following problem, whose optimum value yields an upper bound on the exact objective in (22):

$$ \min_{p(\theta \mid x_P, x_V)} \mathbb{E}_{p(\phi)p(x_P, x_V \mid \phi)p(\theta \mid x_P, x_V)} \left[ \log \frac{p(\theta \mid x_P, x_V)}{q(\theta)} - \beta \log q(x_P \mid \theta) - \gamma \log q(x_V \mid x_P, \theta) \right] $$

$$ = \min_{p(\theta \mid x_P, x_V)} \mathbb{E}_{p(\phi)p(x_P, x_V \mid \phi)} \left\{ \mathbb{E}_{p(\theta \mid x_P, x_V)} \left[ \log \frac{p(\theta \mid x_P, x_V)}{q(\theta)q(x_P \mid \theta)q(x_V \mid x_P, \theta)^{\gamma}} \right] x_P, x_V \right\} $$

$$ = \min_{p(\theta \mid x_P, x_V)} \mathbb{E}_{p(\phi)p(x_P, x_V \mid \phi)} \left[ D_{KL} \left( p(\cdot \mid x_P, x_V) \left\| \frac{q(\cdot)q(x_P \mid \cdot)q(x_V \mid x_P, \cdot)^{\gamma}}{Z_{\beta,\gamma}(x_P, x_V)} \right) \right] \right], $$

(28)
where
\[
Z_{\beta,\gamma}(x_P, x_V) = \sum_{\psi} q(\psi) q(x_P | \psi)^\beta q(x_V | x_P, \psi)^\gamma
\]
or
\[
Z_{\beta,\gamma}(x_P, x_V) = \int q(\psi) q(x_P | \psi)^\beta q(x_V | x_P, \psi)^\gamma \, d\psi
\]
and we emphasize that the expectation in (28) is over the true distributions of $\theta, (x_P, x_V), \phi$. Note that the use of the variational approximation $q(\theta)$ has eliminated the dependence on the distribution of $x_F$.

It follows from (29) that the optimum distribution $p(\theta | x_P, x_V)$ is given by
\[
p(\theta | x_P, x_V) = \frac{q(\theta) q(x_P | \theta)^\beta q(x_V | x_P, \theta)^\gamma}{Z_{\beta,\gamma}(x_P, x_V)}.
\]

Further, from (23) and (24), we can prove the following result in the same way as Theorem 1.

**Theorem 2** If training with the validation set is successful, i.e., we achieve $I(\theta; x_P) > I_0'$ and $I(\theta; x_V | x_P) > I_0''$, then the optimum information processing for (28) is Bayesian inference:
\[
p(\theta | x_P) \propto q(\theta) q(x_P | \theta)^\beta q(x_V | x_P, \theta)^\gamma
\]
where $q(\theta), q(x_V | \theta)$, and $q(x_V | x_P, \theta)$ are as defined in Sec. 4.4.

**References**

[1] A. Zellner, “Optimal Information Processing and Bayes’s Theorem,” *The American Statistician*, vol. 42, no. 4, pp. 278–280, Nov. 1988.

[2] A.A. Alemi, “Variational Predictive Information Bottleneck,” [https://arxiv.org/abs/1910.10831](https://arxiv.org/abs/1910.10831).

[3] Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, “Estimating Information Flow in Deep Neural Networks,” *Proc. 36th Intl. Conf. Mach. Learn.*, Long Beach, California, PMLR 97, 2019, [https://arxiv.org/abs/1810.05728](https://arxiv.org/abs/1810.05728).