A. Notes on Information in Deep Neural Networks

As a supplement to Section 3, we here provide a detailed discussion on information in deep neural networks. Consider a classification task with a conditional distribution $p(\mathrm{y}|\mathbf{x})$, where the random variable $\mathbf{x} \sim p(\mathbf{x})$ denotes an input (*e.g.*, an image) and the random variable $\mathrm{y} \in \mathbb{Y} = \{1, \cdots, C\}$ denotes the target. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ be a training dataset *i.i.d.* sampled from $p(\mathbf{x}, \mathrm{y}) = p(\mathbf{x})p(\mathrm{y}|\mathbf{x})$. A neural network $f(\cdot; \boldsymbol{\theta})$ with the weight parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ encodes a conditional distribution $q_{\boldsymbol{\theta}}(\mathrm{y}|\mathbf{x})$ by fitting the dataset. We denote the network prediction for $\boldsymbol{x}_i$ as $\hat{y}_i$.

The objective is to minimize the negative log likelihood loss (*i.e.*, the cross-entropy loss) on $\mathcal{D}$:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{\theta}) \tag{A.1}$$

$$= \frac{1}{N}\sum_{i=1}^{N} -\log q_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i). \tag{A.2}$$

We take the mini-batch stochastic gradient descent (SGD) for the minimization. Let $\mathcal{B} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_{\mathcal{B}}}$ denotes a random mini batch *i.i.d.* sampled from $\mathcal{D}$. The average loss gradient of the random mini-batch is defined as $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}) = \frac{1}{n_{\mathcal{B}}}\sum_{\boldsymbol{x}_i \in \mathcal{B}} \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{x}_i, y_i; \boldsymbol{\theta})$. Given a learning rate η and time step t, we update the weight parameter by:

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^{t} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}). \tag{A.3}$$

We introduce **KL divergence** (also termed as relative entropy) as a useful notion for the following discussion; it can be seen as an intrinsic dissimilarity metric for distributions (Kullback and Leibler, 1951). For any two distributions $q_{\boldsymbol{\theta}}(\cdot)$ and $q_{\boldsymbol{\theta}'}(\cdot)$, the KL divergence between them is:

$$D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\cdot)\|q_{\boldsymbol{\theta}'}(\cdot)) \triangleq \mathbb{E}_{q_{\boldsymbol{\theta}}(\cdot)}\left[\log\frac{q_{\boldsymbol{\theta}}(\cdot)}{q_{\boldsymbol{\theta}'}(\cdot)}\right]. \tag{A.4}$$

A.1 Fisher Information

In this sub-section, we first introduce Fisher information from its original parameter estimation view (Fisher, 1922). However, we note that this view is unsuitable in the context of deep learning, thus we discuss an alternative view by taking Fisher information as a **local distance metric for distributions**. This will foster the interpretation of Fisher information as a **measure for information retained by the neural network**, and lay a good foundation for understanding the role of Fisher information in the optimization dynamics and generalization.

Definition A.1. (**Fisher information — an estimation view**). *Imagine $\boldsymbol{\theta}$ as an unknown parameter modeling a distribution $q_{\boldsymbol{\theta}}(\cdot)$. Assuming the log likelihood function $\log q_{\boldsymbol{\theta}}(\cdot)$ is differentiable, we define the* **score function** *of $q_{\boldsymbol{\theta}}(\cdot)$ as the gradient of the log likelihood function $\log q_{\boldsymbol{\theta}}(\cdot)$:*

$$s(\boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\cdot). \tag{A.5}$$

The **Fisher information** *of $q_{\boldsymbol{\theta}}(\cdot)$ is defined as the covariance of the score function:*

$$\mathbf{F}(\boldsymbol{\theta}) \triangleq \mathrm{Cov}_{q_{\boldsymbol{\theta}}(\cdot)}\left[s(\boldsymbol{\theta})\right] \tag{A.6}$$

$$\overset{(a)}{=} \mathbb{E}_{q_{\boldsymbol{\theta}}(\cdot)}\left[(\nabla_{\boldsymbol{\theta}}\log q_{\boldsymbol{\theta}}(\cdot))(\nabla_{\boldsymbol{\theta}}\log q_{\boldsymbol{\theta}}(\cdot))^{\top}\right], \tag{A.7}$$

where (a) follows from the fact that $\mathbb{E}_{q_{\boldsymbol{\theta}}(\cdot)}\left[s(\boldsymbol{\theta})\right] = 0$.

Remark A.1. *For neural network parameterized by* $\boldsymbol{\theta}$*,* $\mathbf{F}(\boldsymbol{\theta})$ *is the* fitted gradient covariance *of the negative log likelihood loss* $\ell(\boldsymbol{x}_i, \hat{y}_i; \boldsymbol{\theta})$*, that is* $\mathbf{F}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{x}_i \sim \mathcal{D}, \hat{y}_i \sim q_{\boldsymbol{\theta}}(\mathrm{y}|\mathbf{x})} \left[\nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\hat{y}_i|\boldsymbol{x}_i)\right]$.

As a side note, it is important not to confuse this with the training gradient covariance, *i.e.*, the **empirical Fisher information**, in which $\hat{y}_i$ above is replaced with the true target $y_i \sim p(\mathrm{y}|\mathbf{x})$. The emprical Fisher information may be used to approximate the Fisher information yet it has some limitations (Kunstner et al., 2019), which are not of interests of discussions here.

This classical view is concerned with the parameter estimation problem, and $\boldsymbol{\theta}$ is treated as an *unknown parameter* characterizing a distribution $q_{\boldsymbol{\theta}}(\cdot)$. A high $\mathbf{F}(\boldsymbol{\theta})$ implies that the likelihood function $q_{\boldsymbol{\theta}}(\cdot)$ is rapidly varying,[1] *i.e.*, the likelihood is easily affected by the choice (or value) of $\boldsymbol{\theta}$; thus it is easy to infer the true value of $\boldsymbol{\theta}$ by sampling data from $q_{\boldsymbol{\theta}}(\cdot)$; in other words, the samples can can provide high amount of information to estimate the unknown parameter $\boldsymbol{\theta}$.[2]

However, the estimation view is fundamentally improper to analyze neural networks. In the context of deep learning, $\boldsymbol{\theta}$ is rather a *known parameter* (*i.e.*, network weights) controling the prediction distribution. It is meaningless to estimate $\boldsymbol{\theta}$ through sampling from predictions. The point of focus is not the information that samples from predictions can provide about $\boldsymbol{\theta}$, but rather the information that $\boldsymbol{\theta}$ can provide about the neural network learning process (little literature has explicitly noted this slight distinction, but it can bring potential confusions).

Learning is a dynamic process. The weight parameter $\boldsymbol{\theta}$ is ever changing (updating) with SGD; it is natural to think of the information of $\boldsymbol{\theta}$ from a variation perspective.[3] Intuitively:

> *If small variation in* $\boldsymbol{\theta}$ *results in large discrepancy to the network prediction distribution* $q_{\boldsymbol{\theta}}(\cdot)$*, this* $\boldsymbol{\theta}$ *can be seen as to withhold high amount of information about the learning process.*

We hereby introduce Fisher information as a local metric for such distribution discrepancy.

Lemma A.1. *Assume that the log likelihood function* $\log q_{\boldsymbol{\theta}}(\cdot)$ *is twice differentiable, and denote its Hessian with respect to* $\boldsymbol{\theta}$ *as* $\mathbf{H}_{\log q_{\boldsymbol{\theta}}(\cdot)}$*. The Fisher information of* $q_{\boldsymbol{\theta}}(\cdot)$ *equals the negative of the expected Hessian (i.e., second derivative) of the log likelihood function, that is:*

$$\mathbf{F}(\boldsymbol{\theta}) = -\mathbb{E}_{q_{\boldsymbol{\theta}}(\cdot)}\left[\mathbf{H}_{\log q_{\boldsymbol{\theta}}(\cdot)}\right]. \tag{A.8}$$

Proof. Let's instantiate the likelihood function by $q_{\boldsymbol{\theta}}(\mathbf{z})$, where $\mathbf{z} \sim q_{\boldsymbol{\theta}}(\mathbf{z})$.[4] Then,

$$\begin{aligned}
\mathbf{H}_{\log q_{\boldsymbol{\theta}}(\mathbf{z})} &\triangleq \nabla^2_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z}) \\
&= \nabla_{\boldsymbol{\theta}} \frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\theta}}(\mathbf{z})} \\
&\overset{(a)}{=} \frac{q_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{H}_{q_{\boldsymbol{\theta}}(\mathbf{z})} - \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z}) \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})^\top}{q_{\boldsymbol{\theta}}(\mathbf{z}) q_{\boldsymbol{\theta}}(\mathbf{z})} \\
&= \frac{\mathbf{H}_{q_{\boldsymbol{\theta}}(\mathbf{z})}}{q_{\boldsymbol{\theta}}(\mathbf{z})} - \left(\frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\theta}}(\mathbf{z})}\right)\left(\frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\theta}}(\mathbf{z})}\right)^\top,
\end{aligned} \tag{A.9}$$

where (a) follows from the quotient rule and $\mathbf{H}_{q_{\boldsymbol{\theta}}(\mathbf{z})} \triangleq \nabla^2_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})$.

[1] In this note, $q_{\boldsymbol{\theta}}(\cdot)$ may denote a *distribution* or a *likelihood function*. We will add clarifications when necessary.

[2] Plus, Cramér–Rao Bound (Cramér, 1946) shows the precision of any unbiased estimator for $\boldsymbol{\theta}$ is at most $\mathbf{F}(\boldsymbol{\theta})$.

[3] Notice how this differs from the common definition of information by *Shannon entropy*, which considers information about the *static* distribution per se, instead of the *change* of the distribution by the local parameter variation which we are discussing. An important prior work of this is Achille et al. (2019).

[4] The random variable $\mathbf{z}$ can be viewed as a imprecise shorthand notation for $\mathrm{y}|\mathbf{x}$ in our context.

Taking expectation on Eq. A.9:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\mathbf{H}_{\log q_{\boldsymbol{\theta}}(\mathbf{z})}\right] &= \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\frac{\mathbf{H}_{q_{\boldsymbol{\theta}}(\mathbf{z})}}{q_{\boldsymbol{\theta}}(\mathbf{z})} - \left(\frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\theta}}(\mathbf{z})}\right)\left(\frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\theta}}(\mathbf{z})}\right)^{\top}\right] \\
&= \int \frac{\mathbf{H}_{q_{\boldsymbol{\theta}}(\mathbf{z})}}{q_{\boldsymbol{\theta}}(\mathbf{z})} q_{\boldsymbol{\theta}}(\mathbf{z})\mathrm{d}\mathbf{z} - \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\left(\frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\theta}}(\mathbf{z})}\right)\left(\frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\theta}}(\mathbf{z})}\right)^{\top}\right] \\
&= \int \nabla^2_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})\mathrm{d}\mathbf{z} - \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z}) \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z})^{\top}\right] \\
&\overset{(b)}{=} \nabla^2_{\boldsymbol{\theta}} \int q_{\boldsymbol{\theta}}(\mathbf{z})\mathrm{d}\mathbf{z} - \mathbf{F}(\boldsymbol{\theta}) \\
&= -\mathbf{F}(\boldsymbol{\theta}), \qquad \text{(A.10)}
\end{aligned}
$$

where (b) follows from Definition A.1. Thus $\mathbf{F}(\boldsymbol{\theta}) = -\mathbb{E}_{q_{\boldsymbol{\theta}}(\cdot)}\left[\mathbf{H}_{\log q_{\boldsymbol{\theta}}(\cdot)}\right]$. □

Corollary A.1. (**Fisher information — a metric view**). *Given two distributions $q_{\boldsymbol{\theta}}(\cdot)$ and $q_{\boldsymbol{\theta}'}(\cdot)$ parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ respectively and assuming their likelihood functions are twice differentiable, we have the Fisher information of $\boldsymbol{\theta}$ as the following:*

$$
\mathbf{F}(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}'} D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\cdot)\|q_{\boldsymbol{\theta}'}(\cdot)\right)\Big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}. \qquad \text{(A.11)}
$$

Proof. By definition in Eq. A.4 and instantiating the likelihood function with the random variable $\mathbf{z}$, the gradient of the KL divergence can be decomposed as:

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}'} D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\mathbf{z})\|q_{\boldsymbol{\theta}'}(\mathbf{z})\right) &= \nabla_{\boldsymbol{\theta}'}\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log q_{\boldsymbol{\theta}}(\mathbf{z})\right] - \nabla_{\boldsymbol{\theta}'}\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log q_{\boldsymbol{\theta}'}(\mathbf{z})\right] \\
&= -\nabla_{\boldsymbol{\theta}'}\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log q_{\boldsymbol{\theta}'}(\mathbf{z})\right] \\
&= -\int q_{\boldsymbol{\theta}}(\mathbf{z})\nabla_{\boldsymbol{\theta}'} \log q_{\boldsymbol{\theta}'}(\mathbf{z})\mathrm{d}\mathbf{z}.
\end{aligned}
$$

Thus, the second derivative of the KL divergence with regard to $\boldsymbol{\theta}'$ evaluated at $\boldsymbol{\theta}' = \boldsymbol{\theta}$ is:

$$
\begin{aligned}
\nabla^2_{\boldsymbol{\theta}'} D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\mathbf{z})\|q_{\boldsymbol{\theta}'}(\mathbf{z})\right)\Big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} &= -\int q_{\boldsymbol{\theta}}(\mathbf{z})\nabla^2_{\boldsymbol{\theta}'} \log q_{\boldsymbol{\theta}'}(\mathbf{z})\Big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}\mathrm{d}\mathbf{z} \\
&= -\int q_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{H}_{\log q_{\boldsymbol{\theta}}(\mathbf{z})}\mathrm{d}\mathbf{z} \\
&= -\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\mathbf{H}_{\log q_{\boldsymbol{\theta}}(\mathbf{z})}\right] \\
&\overset{(a)}{=} \mathbf{F}(\boldsymbol{\theta}), \qquad \text{(A.12)}
\end{aligned}
$$

where (a) follows from Lemma A.1. Thus $\mathbf{F}(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}'} D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\cdot)\|q_{\boldsymbol{\theta}'}(\cdot)\right)\Big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$. □

Corollary A.2.. [Fisher information and the 2$^{\text{nd}}$ order approximation of KL divergence] *Given a distribution $q_{\theta}(\cdot)$ parametrized by $\boldsymbol{\theta}$, consider perturbing the parameter by $\boldsymbol{\epsilon}$ such that $\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, the KL divergence[5] of the two distributions is:*

$$
D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\cdot)\|q_{\boldsymbol{\theta}'}(\cdot)\right) = \frac{1}{2}\boldsymbol{\epsilon}^{\top}\mathbf{F}(\boldsymbol{\theta})\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2). \qquad \text{(A.13)}
$$

[5] It is easy to verify that the sign of $\boldsymbol{\epsilon}$ does not affect the result in Eq. A.13; in this regard, Fisher information may be considered as a *symmetric metric* in the distribution space.

Proof. Instantiate the likelihood function with the random variable $\mathbf{z}$. The second order Taylor expansion of $\log q_{\boldsymbol{\theta}'}(\mathbf{z})$ with regard to θ is given as:

$$\log q_{\boldsymbol{\theta}'}(\mathbf{z}) = \log q_{\boldsymbol{\theta}}(\mathbf{z}) + \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z})^\top \boldsymbol{\epsilon} + \frac{1}{2}\boldsymbol{\epsilon}^\top \nabla^2_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z})\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2). \tag{A.14}$$

We then expand the KL divergence between the two distributions by:

$$\begin{aligned}
D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\mathbf{z}) \| q_{\boldsymbol{\theta}'}(\mathbf{z})\right) &= \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log q_{\boldsymbol{\theta}}(\mathbf{z}) - \log q_{\boldsymbol{\theta}'}(\mathbf{z})\right] \\
&\overset{(a)}{=} \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}[\log q_{\boldsymbol{\theta}}(\mathbf{z}) - \log q_{\boldsymbol{\theta}}(\mathbf{z}) - \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z})^\top \boldsymbol{\epsilon} - \frac{1}{2}\boldsymbol{\epsilon}^\top \nabla^2_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z})\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2)] \\
&= -\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z})^\top \boldsymbol{\epsilon}\right] - \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\frac{1}{2}\boldsymbol{\epsilon}^\top \nabla^2_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z})\boldsymbol{\epsilon}\right] + o(\|\boldsymbol{\epsilon}\|_2^2) \\
&= -\int \frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\mathbf{z})^\top}{q_{\boldsymbol{\theta}}(\mathbf{z})} q_{\boldsymbol{\theta}}(\mathbf{z})\boldsymbol{\epsilon}\mathrm{d}\mathbf{z} - \frac{1}{2}\boldsymbol{\epsilon}^\top \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\nabla^2_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\mathbf{z})\right]\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2) \\
&\overset{(b)}{=} -\boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{\theta}} \int q_{\boldsymbol{\theta}}(\mathbf{z})\mathrm{d}\mathbf{z} + \frac{1}{2}\boldsymbol{\epsilon}^\top \mathbf{F}(\boldsymbol{\theta})\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2) \\
&= -\boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{\theta}} \mathbf{1} + \frac{1}{2}\boldsymbol{\epsilon}^\top \mathbf{F}(\boldsymbol{\theta})\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2) \\
&= \frac{1}{2}\boldsymbol{\epsilon}^\top \mathbf{F}(\boldsymbol{\theta})\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2),
\end{aligned} \tag{A.15}$$

where (a) follows from Eq. A.14, and (b) follows from Lemma A.1 that $\mathbf{F}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}(\mathbf{z})}\left[\mathbf{H}_{\log q_{\boldsymbol{\theta}}(\mathbf{z})}\right]$. Thus $D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\cdot)\|q_{\boldsymbol{\theta}'}(\mathbf{z})\right) = \frac{1}{2}\boldsymbol{\epsilon}^\top \mathbf{F}(\boldsymbol{\theta})\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2)$. □

This metric view offers a sensible interpretation on *common* trends of the Fisher information of $q_{\boldsymbol{\theta}}(\mathrm{y}|\mathbf{x})$ during network training. Relatively speaking:

- ***Low*** $\mathbf{F}(\boldsymbol{\theta})$: This implies that the gradient update will not change the prediction distribution much. It usually happens at (1) the *early phases* of training, where the prediction distribution is close to random, and a small variation to the parameter of the random will have little influence on the distribution; or (2) the *ending phases* (or converging phases), where the training is rather stabilized and the prediction distribution are close to true distribution, thus the gradients are close to zero, leading to a low Fisher information (also see Remark A.1).
- ***High*** $\mathbf{F}(\boldsymbol{\theta})$: This implies that even a small perturbation to the parameter can bring large discrepancy of the network prediction. Empirically, this typically happens when the network is (1) *learning new concepts* (*cf.* Achille et al. (2019)), or in other words *at the fitting phase* (in contrast to the generalization phase, *cf.* Shwartz-Ziv and Tishby (2017), especially on the relation to the gradient signal-to-noise ratio), or (2) *memorizing many hard or noisy examples* (*cf.* Jastrzebski et al. (2021)).

Often, $\mathbf{F}(\boldsymbol{\theta})$ will first increase, then drop, leading to a (skewed) bell-shaped trend during training. Thus the learning process appears to be crossing a barrier or a bottleneck.

The advantage of $\mathbf{F}(\boldsymbol{\theta})$ as a metric for the information content of neural networks is that it provides a dynamic (or variation) perspective compatible with the dynamic learning process, and it is relatively easy to compute. However, the limitation is also obvious — it can only serve as a *local* metric.

Next, we discuss other metric to capture the information content in the neural networks, and will offer a general big picture on the connections between them.