1. Problem statement for biased datasets

Using definitions of $q(\boldsymbol{x},\boldsymbol{y})$ and $p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})$, (2) can be analytically derived as

$$\begin{aligned}&D_{KL}(Q_{\boldsymbol{x},\boldsymbol{y}}\|P_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}))=\\&\int\int q(\boldsymbol{y}|\boldsymbol{x})q(\boldsymbol{x})\log\frac{q(\boldsymbol{y}|\boldsymbol{x})q(\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})q(\boldsymbol{x})}d\boldsymbol{y}d\boldsymbol{x}=\\&\int q(\boldsymbol{x})\int q(\boldsymbol{y}|\boldsymbol{x})\log\frac{q(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}d\boldsymbol{y}d\boldsymbol{x}=\\&\mathbb{E}_{Q_{\boldsymbol{x}}}[D_{KL}(Q_{\boldsymbol{y}|\boldsymbol{x}}\|P_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{\theta}))].\end{aligned}$$

Assuming that $Q_{\boldsymbol{y}|\boldsymbol{x}}$ can be replaced by empirical $\hat{Q}_{\boldsymbol{y}|\boldsymbol{x}}$ and $\boldsymbol{y}=\mathbf{1}_d\in\mathbb{R}^D$ is one-hot vector with only dth class not equal to zero, (4) can be derived as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta})&=\frac{1}{N^b}\sum_{i\in\mathbb{N}^b}[D_{KL}(Q_{\boldsymbol{y}_i|\boldsymbol{x}_i}\|P_{\boldsymbol{y}_i|\boldsymbol{x}_i}(\boldsymbol{\theta}))]=\\&=\frac{1}{N^b}\sum_{i\in\mathbb{N}^b}\sum_{d=1}^{D}\mathbf{1}_d(i)\log\frac{\mathbf{1}_d(i)}{p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}=\\&-\frac{1}{N^b}\sum_{i\in\mathbb{N}^b}\log p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}).\end{aligned}$$

2. Relationship between $D_{KL}(P^{\text{v}}_{\boldsymbol{z}}\|P_{\boldsymbol{z}})$ and Fisher information

Using the sufficiency property [1], we approximate our optimal acquisition function (5) using the distributions of learned representations $\boldsymbol{z}$ as

$$\mathcal{R}_{opt}(b,P)=\underset{\mathcal{R}(b,P)}{\arg\min}\,D_{KL}(\hat{P}^{\text{v}}_{\boldsymbol{z}}\|\hat{P}_{\boldsymbol{z}}),$$

Then, a *connection* between the main task (2) and $D_{KL}(P^{\text{v}}_{\boldsymbol{z}}\|P_{\boldsymbol{z}})$ minimization in (7) via Fisher information can be derived with respect to small perturbations in $\boldsymbol{\theta}$. Assuming that the task model minimizes distribution shift in (2) every backward pass as

$$p^{\text{v}}(\boldsymbol{z}|\boldsymbol{\theta})=p(\boldsymbol{z}|\boldsymbol{\theta})+\Delta p,$$

where $\Delta p=\Delta\boldsymbol{\theta}\frac{\partial p(\boldsymbol{z}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ and $\Delta\to 0$.

By substituting (8), the expanded form of $D_{KL}(P^{\text{v}}_{\boldsymbol{z}}\|P_{\boldsymbol{z}})$ can be written as

$$\begin{aligned}&D_{KL}(P^{\text{v}}_{\boldsymbol{z}}\|P_{\boldsymbol{z}})=\int(p(\boldsymbol{z}|\boldsymbol{\theta})+\Delta p)\log\frac{p(\boldsymbol{z}|\boldsymbol{\theta})+\Delta p}{p(\boldsymbol{z}|\boldsymbol{\theta})}d\boldsymbol{z}=\\&\int(p(\boldsymbol{z}|\boldsymbol{\theta})+\Delta p)\log\left(1+\frac{\Delta p}{p(\boldsymbol{z}|\boldsymbol{\theta})}\right)d\boldsymbol{z}.\end{aligned}$$

Using the Taylor series of natural logarithm, this can be approximated by

$$\begin{aligned}&D_{KL}(P^{\text{v}}_{\boldsymbol{z}}\|P_{\boldsymbol{z}})\approx\int(p(\boldsymbol{z}|\boldsymbol{\theta})+\Delta p)\times\\&\left(\frac{\Delta p}{p(\boldsymbol{z}|\boldsymbol{\theta})}-\frac{(\Delta p)^2}{2(p(\boldsymbol{z}|\boldsymbol{\theta}))^2}\right)d\boldsymbol{z}=\int\Delta p d\boldsymbol{z}+\\&\frac{1}{2}\int\left(\frac{\Delta p}{p(\boldsymbol{z}|\boldsymbol{\theta})}\right)^2p(\boldsymbol{z}|\boldsymbol{\theta})d\boldsymbol{z}-\int\frac{(\Delta p)^3}{2p(\boldsymbol{z}|\boldsymbol{\theta})^2}d\boldsymbol{z},\end{aligned}$$

where the first term using the definition of Δp is equal to zero and the third $\mathcal{O}(\Delta\boldsymbol{\theta}^3)\to 0$.

By substituting Δp and rewriting vector $\boldsymbol{\theta}$ as a discrete sum, the term

$$\frac{\Delta p}{p(\boldsymbol{z}|\boldsymbol{\theta})}\approx\sum_i\frac{\partial\log p(\boldsymbol{z}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}\Delta\boldsymbol{\theta}_i.$$

Using this approximation, the final form of (7) can be obtained as

$$\begin{aligned}&\mathcal{R}_{opt}(b,P)=\underset{\mathcal{R}(b,P)}{\arg\min}\,D_{KL}(P^{\text{v}}_{\boldsymbol{z}}\|P_{\boldsymbol{z}})\\&\approx\underset{\mathcal{R}(b,P)}{\arg\min}\sum_{m,n}\mathcal{I}_{m,n}\Delta\boldsymbol{\theta}_m\Delta\boldsymbol{\theta}_n\approx\underset{\mathcal{R}(b,P)}{\arg\min}\,\Delta\boldsymbol{\theta}^T\mathcal{I}\Delta\boldsymbol{\theta},\end{aligned}$$

where $\mathcal{I}=\mathbb{E}_{P_{\boldsymbol{z}}}\left[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})^T\right]$ is a Fisher information matrix and $\boldsymbol{g}(\boldsymbol{\theta})=\frac{\partial\log p(\boldsymbol{z}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ is a Fisher score with respect to $\boldsymbol{\theta}$.

3. Practical Fisher kernel for DNNs

Using the chain rule for a DNN layer ($\tilde{\boldsymbol{z}}_i^j=\boldsymbol{\theta}^T\boldsymbol{z}_i^j=\boldsymbol{\theta}^T\sigma(\tilde{\boldsymbol{z}}_i^{j-1})$) with $\sigma(\cdot)$ nonlinearity, Jacobian of interest can be simplified as follows

$$\frac{\partial L(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)}{\partial\boldsymbol{\theta}}=\frac{\partial L(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)}{\partial\tilde{\boldsymbol{z}}_i}\frac{\partial\tilde{\boldsymbol{z}}_i}{\partial\boldsymbol{\theta}}=\frac{\partial L(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)}{\partial\tilde{\boldsymbol{z}}_i}\boldsymbol{z}_i^T=\boldsymbol{g}_i\boldsymbol{z}_i^T,$$

where $\boldsymbol{\theta}\in\mathbb{R}^{L\times L}$, $\boldsymbol{z}_i\in\mathbb{R}^{L\times 1}$, and $\boldsymbol{g}_i\in\mathbb{R}^{L\times 1}$.

Then, approximation of FK in (11) for $\boldsymbol{g}_i(\boldsymbol{\theta})=\text{vec}(\partial L(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)/\partial\boldsymbol{\theta})\in\mathbb{R}^{L^2\times 1}$ can be derived as

$$\begin{aligned}&R_{z,g}(\boldsymbol{z}_m,\boldsymbol{z}_n)=\boldsymbol{g}_m(\boldsymbol{\theta})^T\mathcal{I}^{-1}\boldsymbol{g}_n(\boldsymbol{\theta})\overset{\text{PFK}}{\approx}\boldsymbol{g}_m(\boldsymbol{\theta})^T\boldsymbol{g}_n(\boldsymbol{\theta})=\\&\text{vec}\left(\frac{\partial L(\boldsymbol{y}_m,\hat{\boldsymbol{y}}_m)}{\partial\tilde{\boldsymbol{z}}_m}\boldsymbol{z}_m^T\right)^T\text{vec}\left(\frac{\partial L(\boldsymbol{y}_n,\hat{\boldsymbol{y}}_n)}{\partial\tilde{\boldsymbol{z}}_n}\boldsymbol{z}_n^T\right)=\\&\text{vec}\left(\boldsymbol{g}_m\boldsymbol{z}_m^T\right)^T\text{vec}\left(\boldsymbol{g}_n\boldsymbol{z}_n^T\right)=[g_m^1\boldsymbol{z}_m,g_m^2\boldsymbol{z}_m,\ldots,g_m^L\boldsymbol{z}_m]^T\times\\&[g_n^1\boldsymbol{z}_n,g_n^2\boldsymbol{z}_n,\ldots,g_n^L\boldsymbol{z}_n]=\boldsymbol{z}_m^T\boldsymbol{z}_n\sum_l^L g_m^l g_n^l=\boldsymbol{z}_m^T\boldsymbol{z}_n\boldsymbol{g}_m^T\boldsymbol{g}_n.\end{aligned}$$

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, pages 1947–1980, 2018.