1. Problem statement for biased datasets

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

\[
D_{KL}(Q_{x,y}||P_{x,y}(\theta)) = \int \int q(y|x)q(x) \log \frac{q(y|x)q(x)}{p(y|x, \theta)q(x)} \, dy \, dx = \int q(x) \int q(y|x) \log \frac{q(y|x)}{p(y|x, \theta)} \, dy \, dx = \mathbb{E}_{Q_x}[D_{KL}(Q_{y|x}||P_{y|x}(\theta))].
\]

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

\[
\mathcal{L}(\theta) = \frac{1}{N_b} \sum_{i \in \mathcal{N}^b} [D_{KL}(Q_{y|x_i}||P_{y|x_i}(\theta))] = \frac{1}{N_b} \sum_{i \in \mathcal{N}^b} \sum_{d=1}^D 1_d(i) \log \frac{1_d(i)}{p(y_i|x_i, \theta)} = - \frac{1}{N_b} \sum_{i \in \mathcal{N}^b} \log p(y_i|x_i, \theta),
\]

2. Relationship between \(D_{KL}(P^\gamma_z||P_z)\) and Fisher information

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

\[
\mathcal{R}_{opt}(b, P) = \underset{\mathcal{R}(b, P)}{\arg \min} D_{KL}(\hat{P}_z^\gamma||\hat{P}_z),
\]

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

\[
p^\gamma(z|\theta) = p(z|\theta) + \Delta p,
\]

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

By substituting (8), the expanded form of \(D_{KL}(P^\gamma_z||P_z)\) can be written as

\[
D_{KL}(P^\gamma_z||P_z) = \int (p(z|\theta) + \Delta p) \log \frac{p(z|\theta) + \Delta p}{p(z|\theta)} \, dz = \int (p(z|\theta) + \Delta p) \log \left(1 + \frac{\Delta p}{p(z|\theta)}\right) \, dz.
\]

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

\[
D_{KL}(P^\gamma_z||P_z) \approx \int (p(z|\theta) + \Delta p) \times
\]

\[
\left(\frac{\Delta p}{p(z|\theta)} - \frac{(\Delta p)^2}{2(p(z|\theta))^2}\right) \, dz = \int \Delta p \, dz + \frac{1}{2} \int \left(\frac{\Delta p}{p(z|\theta)}\right)^2 p(z|\theta) \, dz - \int \frac{(\Delta p)^3}{2p(z|\theta)^3} \, dz,
\]

where the first term using the definition of \(\Delta p\) is equal to zero and the third \(O(\Delta p^3) \to 0\).

By substituting \(\Delta p\) and rewriting vector \(\theta\) as a discrete sum, the term

\[
\frac{\Delta p}{p(z|\theta)} \approx \sum_i \frac{\partial \log p(z|\theta)}{\partial \theta_i} \Delta \theta_i.
\]

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

\[
\mathcal{R}_{opt}(b, P) = \underset{\mathcal{R}(b, P)}{\arg \min} \int \Delta \theta T \mathcal{I} \Delta \theta,
\]

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

3. Practical Fisher kernel for DNNs

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

\[
\frac{\partial L(y_i, \hat{y}_i)}{\partial \theta} = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{z}_i} \frac{\partial \hat{z}_i}{\partial \hat{z}_i} = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{z}_i} z_i^T = g_i z_i^T,
\]

where \(\theta \in \mathbb{R}^{L \times L}, z_i \in \mathbb{R}^{L \times 1}\), and \(g_i \in \mathbb{R}^{L \times 1}\).

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

\[
R_{\theta}(z_m, z_n) = g_m(\theta)^T \mathcal{I}^{-1} g_n(\theta) \approx \text{vec} \left( \frac{\partial L(y_m, \hat{y}_m)}{\partial z_m} \right)^T \text{vec} \left( \frac{\partial L(y_n, \hat{y}_n)}{\partial z_n} \right) = \text{vec} \left( g_m z_m^T \right)^T \text{vec} \left( g_n z_n^T \right) = \left[ g_m^1 z_m, g_m^2 z_m, \ldots, g_m^L z_m \right]^T \times \left[ g_n^1 z_n, g_n^2 z_n, \ldots, g_n^L z_n \right] = z_m^T z_n \sum_l g_m^l g_n^l = z_m^T z_n g_m^T g_n.
\]

References