

1. Derivatives in the distribution matching approach

Given

$$L = C + \text{Tr}\{(L^{tr} + \lambda I)^{-1} \tilde{K} (L^{tr} + \lambda I)^{-1} L^{tr}\} \\ - 2 \text{Tr}\{(L^{tr} + \lambda I)^{-1} \tilde{K}^c (L^{te} + \lambda I)^{-1} K_{X^{te} X^{tr}}\}$$

where $\tilde{K} = K_{Y^{new} Y^{new}}, \tilde{K}^c = K_{Y^{new} Y^{te}}, L^{tr} = K_{X^{tr} X^{tr}}, L^{te} = K_{X^{te} X^{te}}$.

The derivative w.r.t \mathbf{w}, \mathbf{b} is given by:

$$\begin{aligned} \frac{\partial L}{\partial \tilde{K}} &= (K_{X^{tr} X^{tr}} + \lambda I)^{-1} K_{X^{tr} X^{tr}}^\top (K_{X^{tr} X^{tr}} + \lambda I)^{-1} \\ \frac{\partial L}{\partial \tilde{K}_c} &= 2(K_{X^{tr} X^{tr}} + \lambda I)^{-1} K_{X^{te} X^{tr}}^\top (K_{X^{te} X^{te}} + \lambda I)^{-1} \\ \frac{\partial L}{\partial \mathbf{w}_p} &= \text{Tr}[(\frac{\partial L}{\partial \tilde{K}})^\top (\frac{\partial \tilde{K}}{\partial \mathbf{w}_p})] - \text{Tr}[(\frac{\partial L}{\partial \tilde{K}_c})^\top (\frac{\partial \tilde{K}_c}{\partial \mathbf{w}_p})] \\ &= \text{Tr}[(\frac{\partial L}{\partial \tilde{K}})^\top (\tilde{K} \odot D_p)] - \text{Tr}[(\frac{\partial L}{\partial \tilde{K}_c})^\top (\tilde{K}_c \odot E_p)] \\ \frac{\partial L}{\partial \mathbf{b}_p} &= \text{Tr}[(\frac{\partial L}{\partial \tilde{K}})^\top (\frac{\partial \tilde{K}}{\partial \mathbf{b}_p})] - \text{Tr}[(\frac{\partial L}{\partial \tilde{K}_c})^\top (\frac{\partial \tilde{K}_c}{\partial \mathbf{b}_p})] \\ &= \text{Tr}[(\frac{\partial L}{\partial \tilde{K}})^\top (\tilde{K} \odot \tilde{D}_p)] - \text{Tr}[(\frac{\partial L}{\partial \tilde{K}_c})^\top (\tilde{K}_c \odot \tilde{E}_p)] \end{aligned}$$

where

$$\begin{aligned} [D_p]_{ij} &= -\frac{1}{\sigma^2} (Y_i^{new} - Y_j^{new})(Y_i^{tr} I(i=p) - Y_j^{tr} I(j=p)) \\ [E_p]_{ij} &= -\frac{1}{\sigma^2} (Y_i^{new} - Y_j^{te}) Y_i^{tr} I(i=p) \\ [\tilde{D}_p]_{ij} &= -\frac{1}{\sigma^2} (Y_i^{new} - Y_j^{new})(I(i=p) - I(j=p)) \\ [\tilde{E}_p]_{ij} &= -\frac{1}{\sigma^2} (Y_i^{new} - Y_j^{te}) I(i=p) \end{aligned}$$

After parameterizing \mathbf{w}, \mathbf{b} to ensure the smoothness by $\mathbf{w} = R\mathbf{g}, \mathbf{b} = R\mathbf{h}$, where $R = L^{tr}(L^{tr} + \lambda I)^{-1}$. The new parametrization results in:

$$\begin{aligned} [D_p]_{ij} &= -\frac{1}{\sigma^2} (Y_i^{new} - Y_j^{new})(Y_i^{tr} R_{ip} - Y_j^{tr} R_{jp}) \\ [E_p]_{ij} &= -\frac{1}{\sigma^2} (Y_i^{new} - Y_j^{te}) Y_i^{tr} R_{ip} \\ [\tilde{D}_p]_{ij} &= -\frac{1}{\sigma^2} (Y_i^{new} - Y_j^{new})(R_{ip} - R_{jp}) \\ [\tilde{E}_p]_{ij} &= -\frac{1}{\sigma^2} (Y_i^{new} - Y_j^{te}) R_{ip} \end{aligned}$$

2. Derivation of the covariance matrix in Active Learning

Detailed derivation for covariance matrix Σ_2 for $P(Y^{teU} | X^{teU}, X^{tr}, Y^{tr}, X^{teL}, Y^{teL})$.

$$\begin{aligned} &P(Y^{teU} | X^{teU}, X^{tr}, Y^{tr}, X^{teL}, Y^{teL}) \\ &= \int_{Y^{new}} P(Y^{teU}, Y^{new} | X^{teU}, X^{tr}, Y^{tr}, X^{teL}, Y^{teL}) dY^{new} \\ &= \int_{Y^{new}} P(Y^{teU} | X^{teU}, X^{tr}, Y^{new}, X^{teL}, Y^{teL}) \\ &\quad P(Y^{new} | X^{tr}, Y^{tr}, X^{teL}, Y^{teL}) dY^{new} \\ &= C \int_{Y^{new}} \exp\{-\frac{1}{2}(Y^{teU} - \mu)^\top \Sigma^{-1}(Y^{teU} - \mu)\} \\ &\quad \exp\{-\frac{1}{2}(Y^{new} - Y^{tr} - \mu_1)^\top \Sigma_1^{-1}(Y^{new} - Y^{tr} - \mu_1)\} dY^{new} \\ &= C \int_{Y^{new}} \exp\{-\frac{1}{2}(Y_* - \Omega_1 Y^{new})^\top \Sigma^{-1}(Y_* - \Omega_1 Y^{new})\} \\ &\quad \exp\{-\frac{1}{2}(Y^{new} - Y^{tr} - \mu_1)^\top \Sigma_1^{-1}(Y^{new} - Y^{tr} - \mu_1)\} dY^{new} \\ &= C' \int_{Y^{new}} \exp\{-\frac{1}{2}(Y_* - \Omega_1 Y^{new})^\top \Sigma^{-1}(Y_* - \Omega_1 Y^{new})\} \\ &\quad \exp\{-\frac{1}{2}(Y^{new} - \mu_1)^\top \Sigma_1^{-1}(Y^{new} - \mu_1)\} dY^{new} \end{aligned}$$

where $Y_* = Y^{teU} - \Omega_2 Y^{teL}$.

Combining the terms related to Y^{new} , we would have

$$\begin{aligned} &\int_{Y^{new}} \exp\{-\frac{1}{2}[(Y^{new})^\top (\Omega_1^\top \Sigma^{-1} \Omega_1 + \Sigma_1^{-1}) Y^{new} \\ &\quad - 2(Y_*^\top \Sigma^{-1} \Omega_1 + \mu_1^\top \Sigma_1^{-1}) Y^{new}]\} dY^{new} \\ &= C' \exp\{\frac{1}{2}(Y_*^\top \Sigma^{-1} \Omega_1 + \mu_1^\top \Sigma_1^{-1})(\Omega_1^\top \Sigma^{-1} \Omega_1 + \Sigma_1^{-1})^{-1} \\ &\quad (Y_*^\top \Sigma^{-1} \Omega_1 + \mu_1^\top \Sigma_1^{-1})^\top\} \end{aligned}$$

Combining the term related to Y_* , we have Y_* also following a gaussian distribution with the quadratic term as the following:

$$Y_*^\top (\Sigma^{-1} - \Sigma^{-1} \Omega_1 (\Omega_1^\top \Sigma^{-1} \Omega_1 + \Sigma_1^{-1})^{-1} \Omega_1^\top \Sigma^{-1}) Y_*$$

which implies that Y_* , or Y^{teU} has a covariance matrix of:

$$\begin{aligned} \Sigma_2 &= (\Sigma^{-1} - \Sigma^{-1} \Omega_1 (\Omega_1^\top \Sigma^{-1} \Omega_1 + \Sigma_1^{-1})^{-1} \Omega_1^\top \Sigma^{-1})^{-1} \\ &= \Sigma + \Omega_1 \Sigma_1 \Omega_1^\top \quad (\text{matrix inversion lemma}) \\ &= \Sigma + \Omega_1 (\Sigma_0 + K_1 \Sigma_s K_1^\top) \Omega_1^\top \end{aligned}$$