A. Learning kernels

We consider the scenario of learning kernel.

1. Prove the equivalence of the (primal) optimization problem

\[
\min_{w, \mu \in \Delta_q} \frac{1}{2} \sum_{k=1}^{p} \frac{\|w_k\|_2^2}{\mu_k} + C \sum_{i=1}^{m} \max \left\{ 0, 1 - y_i \left( \sum_{k=1}^{p} w_k \cdot \Phi_k(x_i) \right) \right\}.
\]

and the (dual) problem

\[
\max_{\alpha} 2\alpha^\top 1 - \left\| \alpha^\top Y^\top K_1 Y \alpha \right\|_r
\]

subject to: \(0 \leq \alpha \leq C \land \alpha^\top y = 0,\)

where \(r, q \geq 1\) are conjugate numbers, \(\frac{1}{r} + \frac{1}{q} = 1.\)

2. What does the problem correspond to for \(r = 1?\)

B. Deep boosting

Let \(F\) be the function defined over \(\mathcal{F} = \text{conv}(\bigcup_{k=1}^{p} H_k)\) by

\[
F(f) = \hat{R}_{S, \rho}(f) + \frac{4}{\rho} \sum_{t=1}^{T} \alpha_t \mathcal{R}_m(H_{kt}),
\]

for any \(f = \sum_{t=1}^{T} \alpha_t h_t \in \mathcal{F}.\) Define the Voted Risk Minimization (VRM) solution as the function \(f^*\) minimizing \(F:\)

\[
f_{\text{VRM}} = \arg\min_{f \in \mathcal{F}} F(f).
\]
Let \( f^* \) be the element in \( \mathcal{F} \) with the smallest generalization error:
\[
R(f^*) = \inf_{f \in \mathcal{F}} R(f).
\]

1. Fix \( \rho > 0 \). Use the margin bound presented in class for deep boosting to derive an upper bound on \( R(f_{VRM}) - R_\rho(f^*) \), where \( R_\rho(f^*) = E(x,y) \sim D \mathbb{I}[y f^*(x) \leq \rho] \) is the \( \rho \)-margin loss of \( f^* \).

2. Compare this result to generalization bound proven in class for SRM.

C. Structured prediction

1. Show that \( \Phi_u: v \mapsto e^{u-v} \) upper bounds \( v \mapsto u \mathbb{1}_{v \leq 0} \) for all \( u \geq 0 \).

2. Use that to derive a new structured prediction algorithm based on the hypothesis set
\[
\mathcal{H}_2 = \{ x \mapsto w \cdot \Psi(x,y) : w \in \mathbb{R}^N, \|w\|_2 \leq \Lambda_2 \},
\]
for a feature vector \( \Psi \).

3. Assume a bigram feature decomposition \( \Psi(x,y) = \sum_{k=1}^l \phi(x,k,y_{k-1},y_k) \). Use that to give an explicit margin bound, assuming that \( \|\Psi\| \leq r \).

4. Describe in detail an efficient algorithm for the the computation of the gradient for your algorithm.