

(b) We will need to apply the following (in the right order):

$$\forall h \in H, |\varepsilon_\tau(h) - \hat{\varepsilon}_\tau(h)| \leq \bar{\gamma} \quad \text{w.p.}(1 - \delta), \quad \delta = 2K \exp(-2\bar{\gamma}^2 m) \quad (6)$$

$$\varepsilon_\tau = (1 - 2\tau)\varepsilon + \tau, \quad \varepsilon_0 = \frac{\varepsilon_\tau - \tau}{1 - 2\tau} \quad (7)$$

$$\forall h \in H, \hat{\varepsilon}_\tau(\hat{h}) \leq \hat{\varepsilon}_\tau(h), \quad \text{in particular for } h^* \quad (8)$$

Here is the derivation:

$$\varepsilon_0(\hat{h}) = \frac{\varepsilon_\tau(\hat{h}) - \tau}{1 - 2\tau} \quad (9)$$

$$\leq \frac{\hat{\varepsilon}_\tau(\hat{h}) + \bar{\gamma} - \tau}{1 - 2\tau} \quad \text{w.p.}(1 - \delta) \quad (10)$$

$$\leq \frac{\hat{\varepsilon}_\tau(h^*) + \bar{\gamma} - \tau}{1 - 2\tau} \quad \text{w.p.}(1 - \delta) \quad (11)$$

$$\leq \frac{\varepsilon_\tau(h^*) + 2\bar{\gamma} - \tau}{1 - 2\tau} \quad \text{w.p.}(1 - \delta) \quad (12)$$

$$= \frac{(1 - 2\tau)\varepsilon_0(h^*) + \tau + 2\bar{\gamma} - \tau}{1 - 2\tau} \quad \text{w.p.}(1 - \delta) \quad (13)$$

$$= \varepsilon_0(h^*) + \frac{2\bar{\gamma}}{1 - 2\tau} \quad \text{w.p.}(1 - \delta) \quad (14)$$

$$= \varepsilon_0(h^*) + 2\gamma \quad \text{w.p.}(1 - \delta) \quad (15)$$

Where we used in the following order: (7) (6) (8) (6) (7), and the last 2 steps are algebraic simplifications, and defining γ as a function of $\bar{\gamma}$. Now we can fill out $\bar{\gamma} = \gamma(1 - 2\tau)$ into δ of (6), solve for m and we are done.

Note: one could shorten the above derivation and go straight from (9) to (12) by using that result from class.

- (c) The closer τ is to 0.5, the more samples are needed to get the same generalization error bound. For τ approaching 0.5, the training data becomes more and more random; having no information at all about the underlying distribution for $\tau = 0.5$.

6. [19 points] Boosting and high energy physics

Consider the following data set from the [Fermi Photon Reconstruction](https://www.kaggle.com/competitions/fermi-photon-reconstruction) competition. The data set contains information about the energy and direction of photons produced in the collision of two particles. The data is divided into training and testing sets. The training set is used to train a model to predict the energy and direction of photons. The testing set is used to evaluate the model's performance. The data set is available at <https://www.kaggle.com/competitions/fermi-photon-reconstruction>. The data set is a CSV file with 10 columns: 'x', 'y', 'z', 'x2', 'y2', 'z2', 'x3', 'y3', 'z3', and 'x4'. The first three columns represent the coordinates of the photon, and the last four columns represent the coordinates of the collision point. The data set is divided into training and testing sets. The training set is used to train a model to predict the energy and direction of photons. The testing set is used to evaluate the model's performance. The data set is available at <https://www.kaggle.com/competitions/fermi-photon-reconstruction>.