TUM EI 70360: Machine Learning and Optimization
Fall 2023
Lecturer: Reinhard Heckel
Teaching Assistant: Tobit Klug

Problem Set 9
Issued: Tuesday, Dec. 12, 2023
Due: Thursday, Dec. 21, 2023

Problem 1. Let $\mathcal{H}$ be a finite hypothesis class. Let $\hat{h}$ be the empirical risk minimizer and $h_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} R(h)$. Suppose $\sup_{h\in\mathcal{H}} |\hat{R}(h) - R(h)|$ decays as shown by the blue curve in Figure 1 (a), when n increases, where $\hat{R}(h)$ is the empirical risk on n i.i.d. examples and $R(h)$ is the population (true) risk.

i) Suppose $R(h_{\mathcal{H}}) = 2\epsilon$. Draw one possible curve that shows how the population risk $R(\hat{h})$ of $\hat{h}$ changes as a function of n and mark the intervals where the curve locates in at n_1 and n_2.

ii) Suppose $R(h_{\mathcal{H}}) = 0$. Draw one possible curve that shows how the empirical risk $\hat{R}(\hat{h})$ of $\hat{h}$ changes as a function of n.

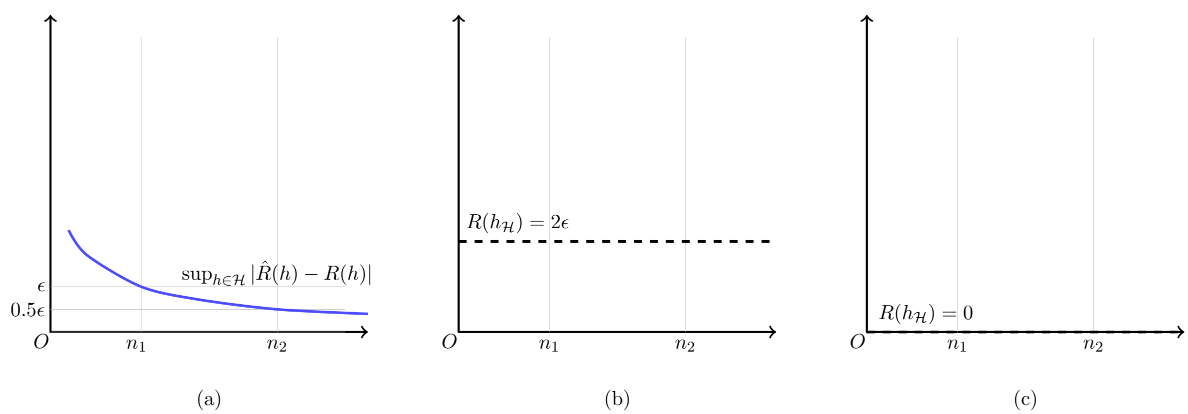

Figure 1: Empirical risk minimization.

Problem 2 (Generalization of SGD iterate)**.** Consider a binary classification problem with a linear classifier parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$. We consider the hinge loss

$$\ell(\boldsymbol{\theta}, \mathbf{x}, y) = \max\{0, 1 - y\langle \boldsymbol{\theta}, \mathbf{x}\rangle\}.$$

Define the empirical risk on n i.i.d. examples and the population risk of $\boldsymbol{\theta}$ as

$$\hat{R}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i), \qquad R(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)}\left[\ell(\boldsymbol{\theta}, \mathbf{x}, y)\right].$$

The empirical risk minimizer is defined as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \hat{R}(\boldsymbol{\theta}).$$

The number of examples n is large, so we apply SGD to minimize $\hat{R}(\boldsymbol{\theta})$ using the update rule

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \alpha_k G(\boldsymbol{\theta}^k).$$

where α_k is the stepsize, and $G(\boldsymbol{\theta})$ is equal to the sub-gradient of $\ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i)$ with respect to $\boldsymbol{\theta}$ where i is chosen uniformly at random from the training examples for each iteration.

1. (3 points) Show that the empirical risk $\hat{R}(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$.

2. (3 points) Compute the sub-gradient $\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}_i, y_i)$ of $\ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i)$ and show that

$$\mathbb{E}\left[G(\boldsymbol{\theta})\right] = \nabla_{\boldsymbol{\theta}} \hat{R}(\boldsymbol{\theta}).$$

3. (3 points) Let $\boldsymbol{\theta}^n$ be the iterate after n SGD iterations. We bound the risk of $\boldsymbol{\theta}^n$. First, it can be shown that with probability at least $1 - \delta_1$ with respect to the SGD updates,

$$\hat{R}(\boldsymbol{\theta}^n) \leq \epsilon_1(n, \delta_1),$$

for some function $\epsilon_1(n, \delta_1)$ of n and δ_1. Second, suppose it can be shown that with probability at least $1 - \delta_2$ with respect to the random draw of the training set,

$$\sup_{\boldsymbol{\theta}} \left(R(\boldsymbol{\theta}) - \hat{R}(\boldsymbol{\theta}) \right) \leq \epsilon_2(n, \delta_2),$$

for some function $\epsilon_2(n, \delta_2)$ of n and δ_2. Show that with probability at least $1 - \delta_1 - \delta_2$ with respect to the SGD updates and the random draw of the training set,

$$R(\boldsymbol{\theta}^n) \leq \epsilon_1(n, \delta_1) + \epsilon_2(n, \delta_2).$$