Foundations of Deep Learning **Fall 2024**

Homework 3: Generalization, Regularization and Adversarial Examples

Lecturer: Aurelien Lucchi

The points of the best-two-out-of-three homeworks, including this one, will be contributed to the final score. The points of each problem in this exercise sheet are equally weighted. Period: 14 November 2024 18:00 - 19 December 2024 23:55 (Bern time).

Problem 1 (PAC Bayes Bounds) (10 Points):

We consider a supervised learning scenario with a hypothesis space $\mathcal{H}$ and a dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d. from an unknown distribution D over $\mathcal{X} \times \mathcal{Y}$. The performance of a hypothesis $h \in \mathcal{H}$ is measured using a loss function $\ell(h, (x, y))$, which quantifies the error of the hypothesis on a given data point (x, y).

The **true risk** $\mathcal{R}(h)$ and the **empirical risk** $\hat{\mathcal{R}}(h)$ of a hypothesis h are defined as:

$$\mathcal{R}(h) = \mathbb{E}_{(x,y)\sim D}[\ell(h, (x, y))], \quad \hat{\mathcal{R}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, (x_i, y_i)).$$

In the PAC-Bayes framework, we extend these definitions to posterior distributions over hypotheses. Let P be a prior distribution over $\mathcal{H}$ and Q a posterior distribution after observing the data. The **true risk** $\mathcal{R}(Q)$ and the **empirical risk** $\hat{\mathcal{R}}(Q)$ of a posterior distribution Q are given by:

$$\mathcal{R}(Q) = \mathbb{E}_{h\sim Q}[\mathcal{R}(h)], \quad \hat{\mathcal{R}}(Q) = \mathbb{E}_{h\sim Q}[\hat{\mathcal{R}}(h)].$$

The **Kullback-Leibler (KL) divergence** between the posterior Q and the prior P is defined as:

$$\mathrm{KL}(Q||P) = \int_{\mathcal{H}} Q(h) \log\left(\frac{Q(h)}{P(h)}\right) dh.$$

The PAC-Bayesian bound provides a high-probability guarantee on the true risk of a hypothesis sampled from the posterior distribution. Here we present another version of the PAC-Bayesian bound from Catoni (2003):

Theorem 1. *For any numbers $\lambda > 0$, $\delta \in (0, 1)$, any distribution Q over $\mathcal{H}$, with probability at least $1 - \delta$ over the choice of the training set S, there exists a constant $C > 0$ independent to λ, δ, Q, S such that:*

$$\mathcal{R}(Q) \leq \hat{\mathcal{R}}(Q) + \frac{\lambda C^2}{8n} + \frac{1}{\lambda}\left(KL(Q||P) + \log\left(\frac{1}{\delta}\right)\right).$$

Here, λ controls the trade-off between the empirical risk and the complexity term involving the KL divergence.

a) Assume that the hypothesis set $\mathcal{H}$ is finite with size $|N|$ and the prior P is the uniform distribution over $\mathcal{H} = \{h_1, ..., h_N\}$ and the posterior Q is in the set of the Dirac masses on $\mathcal{H}$.

i) Show that with probability at least $1 - \delta$ over the choice of the training set S, it holds that

$$\mathcal{R}(Q) \leq \hat{\mathcal{R}}(Q) + \frac{\lambda C^2}{8n} + \frac{1}{\lambda}\left(\log\left(\frac{N}{\delta}\right)\right).$$

ii) By optimizing λ, show that with probability at least $1 - \delta$ over the choice of the training set S, it holds that

$$\mathcal{R}(Q) \leq \hat{\mathcal{R}}(Q) + C\sqrt{\frac{\log \frac{N}{\delta}}{2n}}.$$

Note that we have recovered the uniform Hoeffding bound in the lecture note.

b) Assume that the prior P is a Gaussian distribution $\mathcal{N}(\mu_0, \sigma_0^2 I)$ and the posterior Q is Gaussian $\mathcal{N}(\mu, \sigma^2 I)$ over the parameter space $\mathcal{H} = \mathbb{R}^d$.

i) Using the formula for KL divergence between two Gaussians:

$$\mathrm{KL}(\mathcal{N}(\mu, \Sigma)\|\mathcal{N}(\mu_0, \Sigma_0)) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_0^{-1}\Sigma) + (\mu - \mu_0)^\top \Sigma_0^{-1}(\mu - \mu_0) - d + \log\frac{\det \Sigma_0}{\det \Sigma}\right).$$

to show that with probability at least $1 - \delta$,

$$\mathcal{R}(Q) \le \hat{\mathcal{R}}(Q) + \frac{1}{\lambda}\left(\frac{1}{2}\left(\frac{\|\mu - \mu_0\|_2^2}{\sigma_0^2} + d\left(\frac{\sigma^2}{\sigma_0^2} - 1 - \log\frac{\sigma^2}{\sigma_0^2}\right)\right) + \log\left(\frac{1}{\delta}\right)\right) + \frac{\lambda C^2}{8n}.$$

ii) By optimizing λ, show that with probability at least $1 - \delta$ over the choice of the training set S, it holds that

$$\mathcal{R}(Q) \le \hat{\mathcal{R}}(Q) + C\sqrt{\frac{1}{2n}\left(\frac{1}{2}\left(\frac{\|\mu - \mu_0\|_2^2}{\sigma_0^2} + d\left(\frac{\sigma^2}{\sigma_0^2} - 1 - \log\frac{\sigma^2}{\sigma_0^2}\right)\right) + \log\left(\frac{1}{\delta}\right)\right)}.$$

Problem 2 (Regularization in Linear Regression) (10 Points):
Consider a linear regression problem where the input $\mathbf{x} \in \mathbb{R}^d$ is drawn i.i.d. from a standard isotropic Gaussian distribution $\mathcal{N}(0, I_d)$. The true target coefficient is $\mathbf{w}^* \in \mathbb{R}^d$, and the observed label $y \in \mathbb{R}$ is generated as:

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is centered Gaussian noise with variance σ^2. Given a dataset $(\mathbf{X}, \mathbf{y})$ of n i.i.d. samples, we aim to explore the effects of regularization on the generalization properties of the linear regression estimator.

The **true risk** $\mathcal{R}(\mathbf{w})$ and the **empirical risk** $\hat{\mathcal{R}}(\mathbf{w})$ of a hypothesis $\mathbf{w}$ are defined as follows:

$$\mathcal{R}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)}[(y - \mathbf{x}^\top \mathbf{w})^2], \quad \hat{\mathcal{R}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

The regularized least-squares estimator $\hat{\mathbf{w}}$ with regularization parameter $\lambda > 0$ aims to minimize the regularized empirical risk:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left(\hat{\mathcal{R}}(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2\right) = \arg\min_{\mathbf{w}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the data matrix, $\mathbf{y} \in \mathbb{R}^n$ is the vector of observed labels, $\lambda > 0$ is the regularization parameter, and I_d is the $d \times d$ identity matrix.

a) Write down the closed-form solution of $\hat{\mathbf{w}}$.

b) Substituting $y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$, show that:

$$\mathbb{E}_\epsilon[\mathcal{R}(\hat{\mathbf{w}})] = \mathbb{E}_\epsilon\left[\|\mathbf{w}^* - \hat{\mathbf{w}}\|^2\right] + \sigma^2.$$

c) Show that the expected value can be decomposed as:

$$\mathbb{E}_\epsilon\left[\|\mathbf{w}^* - \hat{\mathbf{w}}\|^2\right] = \text{Bias}^2 + \text{Variance},$$

where

$$\text{Bias}^2 = \lambda^2 \left\| \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda I_d\right)^{-1} \mathbf{w}^* \right\|^2, \qquad \text{Variance} = \sigma^2 \text{Tr}\left((\mathbf{X}^\top\mathbf{X} + n\lambda I_d)^{-2}\mathbf{X}^\top\mathbf{X}\right).$$

Discuss the behavior of these two terms when $\lambda \to 0$ and $\lambda \to \infty$.

d) Note that for $n \gg d$, by the Law of Large Numbers, the covariance matrix $\mathbf{X}^\top\mathbf{X} \approx nI_d$. Substituting this approximation, we obtain the proxies B and V for the two terms:

$$\text{Bias}^2 \approx B^2 = \frac{\lambda^2}{(1+\lambda)^2}\|\mathbf{w}^*\|^2, \qquad \text{Variance} \approx V = \frac{d\sigma^2}{n(1+\lambda)^2}.$$

Now find

$$\lambda^* = \underset{\lambda}{\text{argmin}}(B^2 + V).$$

What can you interpret from the result?

Problem 3 (Adversarial Training in Linear Regression) (10 Points):
Recall the definition of the adversarial loss in adversarial training:

$$R_{\text{adv}}(\mathbf{w}) = R_{\text{adv}}(f_{\mathbf{w}}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\max_{\|\boldsymbol{\delta}\|\leq\rho} \ell(f_{\mathbf{w}}(\mathbf{x}+\boldsymbol{\delta}), y)\right]$$

where $\Delta(\mathbf{x})$ is neighborhood of $\mathbf{x}$. Now we try to write down its analytic form in the case of linear regression: assume that $\mathbf{x} \in \mathcal{N}(0, \mathbf{I}_d)$, $y \sim \mathbf{x}^\top \mathbf{w}^* + \epsilon$ for some fixed $\mathbf{w}^* \in \mathbb{R}^d$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $f_{\mathbf{w}}(\cdot) = (\cdot)^\top \mathbf{w}$ and $\ell(a, b) = (a - b)^2$.

a) Show that
$$R_{\text{adv}}(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^*\|^2 + \sigma^2 + \rho^2\|\mathbf{w}\|^2 + 2\rho c_0\|\mathbf{w}\|\sqrt{\|\mathbf{w} - \mathbf{w}^*\|^2 + \sigma^2}$$
where $c_0 = \sqrt{2/\pi}$.

b) Show that R_{adv} is convex wrt $\mathbf{w}$ and there exists some constant $c > 0$ such that $\mathbf{w}^*_{\text{adv}} := \operatorname{argmin}_{\mathbf{w}} R_{\text{adv}}(\mathbf{w})$ is equal to 0 whenever $\rho \geq c$.

References

Olivier Catoni. A pac-bayesian approach to adaptive classification. *preprint*, 840(2):6, 2003.