TTIC 31230 Fundamentals of Deep Learning

Regularization and Generalization Problems

PAC-Bayes Background Consider any probability distribution $P(h)$ over a discrete class $\mathcal{H}$. Assume $0 \leq \mathcal{L}(h,x,y) \leq L_{\max}$. Define

$$\begin{aligned}\mathcal{L}(h) &= E_{(x,y)\sim\mathrm{Pop}}\;\mathcal{L}(h,x,y)\\ \hat{\mathcal{L}}(h) &= E_{(x,y)\sim\mathrm{Train}}\;\mathcal{L}(h,x,y)\end{aligned}$$

We now have the theorem that with probability at least $1-\delta$ over the draw of training data the following holds simultaneously for all h.

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N}\left(\ln\frac{1}{P(h)} + \ln\frac{1}{\delta}\right)\right) \quad (1)$$

This motivates

$$h^* = \operatorname*{argmin}_h\; \hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N_{\mathrm{train}}}\ln\frac{1}{P(h)} \quad (2)$$

The Bayesian maximum a-posteriori (MAP) rule is

$$h^* = \operatorname*{argmax}_h\; P(h)\prod_{(x,y)\in\mathrm{Train}} P(y|x,h) \quad (3)$$

Problem 1. The Meaning of a PAC-Bayes Prior. Consider an optimal hypothysis for the population distribution.

$$h^* = \operatorname*{argmin}_h\; E_{\langle x,\,y\rangle\sim\mathrm{Pop}}\;\mathcal{L}(h,x,y)$$

Equation (1) holds for any prior P. Consider two priors P_{lucky} and P_{unlucky} where we have

$$P_{\mathrm{lucky}}(h^*) >> P_{\mathrm{unlucky}}(h^*)$$

Explain how equation (1) can hold for both of these priors.

Solution: The prior P should be interpreted as saying which hyopthesis will be measured accurately first as N_{train} increases. We can interpret P as a "guess" as to where we think the good hypotheses are. The prior P is not stating any actual propability of where the optimal hypothesis is. We get accurate measurements first for the hypotheses h for which $P(h)$ is large. For P_{unlucky} we get an accurate measureent of $\mathcal{L}(h^*)$ only much later than we do under P_{lucky}.

Problem 2. Code Length as Probability. Assume that a model h is represented by a (compressed) file $|h|$ bits long. Files have a specific length and no file is a proper prefix of any other file. We say that the set of file bit strings is **prefix free**.

(a) Show that for any prefix-free representation of files as bit strings we have the following Kraft inequality where the sum is over all possible files (of unbounded size).

$$\sum_h 2^{-|h|} \leq 1$$

Solution: Consider a probabilistic process which flips an unbiased coin to determine a next bit until it generates a legal file bit string at which point it outputs that file. This process generates file h with probability $2^{-|h|}$ and, by the prefix-free property, all files can be generated by this process. However, it is possible that this process never terminates The Kraft inequality then follows from $\sum_h P(h) \leq 1$ where we also have $P(\text{divergence}) + \sum_h \; P(h) = 1$.

(b) rewrite (1) in terms of $|h|$ where we take $P(h) = 2^{-|h|}$.

Solution:

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N_{\text{train}}}\left((\ln 2)|h| + \ln\frac{1}{\delta}\right)\right) \quad (1)$$

Problem 3. Comparing Bayesian MAP to PAC-Bayes For $\mathcal{L}(h,x,y) = -\ln P(y|x,h)$ (cross entropy loss) rewrite (2) so as to be as similar to (3) as possible. Note that (1) holds independent of any "truth" of the "prior" P.

Solution:

$$\begin{aligned}
& \underset{h}{\text{argmin}} \left(\frac{1}{N}\sum_{(x,y)\sim\text{Train}} -\ln P(y|x,h)\right) + \frac{5L_{\max}}{N}\ln\frac{1}{P(h)} \\
= \; & \underset{h}{\text{argmax}} \left(\frac{1}{N}\sum_{(x,y)\sim\text{Train}} \ln P(y|x,h)\right) + \frac{5L_{\max}}{N}\ln P(h) \\
= \; & \underset{h}{\text{argmax}} \left(\sum_{(x,y)\sim\text{Train}} \ln P(y|x,h)\right) + 5L_{\max}\ln P(h) \\
= \; & \underset{h}{\text{argmax}} \ln\left(P(h)^{5L_{\max}} \prod_{(x,y)\sim\text{Train}} P(y|x,h)\right) \\
= \; & \underset{h}{\text{argmax}} \; P(h)^{5L_{\max}} \prod_{(x,y)\sim\text{Train}} P(y|x,h)
\end{aligned}$$

Problem 4. Finite Precision Parameters.

(a) Consider a model where the parameter vector Φ has d parameters each of which is represented by a 16 bit floating point number. Express the bound (1) in terms of the dimension d assuming all parameter vectors are equally likely.

Solution:

$$\mathcal{L}(\Phi) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(\Phi) + \frac{5L_{\max}}{N}\left(16d\ln 2 + \ln\frac{1}{\delta}\right)\right)$$

(b) Assume a variable precision representation of numbers where $\Phi[i]$ is given with $|\Phi[i]|$ bits. Express the bound (1) as a function of Φ assuming that $P(\Phi)$ is defined so that each parameter is selected independently and that

$$P(\Phi[i]) = 2^{-|\Phi[i]|}$$

.

Solution:

$$\begin{aligned}\mathcal{L}(h) &\leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N_{\text{train}}}\left(\ln 2|\Phi| + \ln\frac{1}{\delta}\right)\right)\\ |\Phi| &= \sum_i |\Phi[i]|\end{aligned}$$

(c) Repeat part (a) but for a model with I parameters represented by $\Phi[i] = \Psi[k[i]]$ where $k[i]$ is an integer index with $0 \leq k[i] \leq K-1$ and where $\Psi[k]$ is a b-bit floating point number. We define a prior probability on models by selecting each $k[i]$ uniformily from the integers from 0 to $K-1$ and selecting $\Psi[k]$ uniformly from all b-fit floating point numbers.

Solution:

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N}\left(Kb\ln 2 + I\ln k + \ln\frac{1}{\delta}\right)\right)$$

For $I >> K$ this is a much tighter bound than using floating point or even integer representations of parameters. It is a much more compact representaiton of the parameters.

Problem 5. Implicit Bias for SGD on Least Squares Regression. Consider a hypothesis space $\mathcal{H}$ and a learning algorithm $\mathcal{A}$ that maps trainging data to a hyothesis in $\mathcal{H}$. Write $\mathcal{A}(\text{Train})$ for the result of running algorithm $\mathcal{A}$ on training data Train. Also consider a given population distribution Pop where Train consists of N_{train} samples drawn independently from Pop. Let $P_{\mathcal{A},\text{Pop}}(h)$ be the probability that $\mathcal{A}(\text{Train}) = h$ when Train is drawn at random from Pop. The propbability distribution $P_{\mathcal{A},\text{Pop}}$ is independent of any particular training sample and can be used as a PAC-Bayes prior on $\mathcal{H}$. A PAC-Bayes prior represents a learning bias. The distribution $P_{\mathcal{A},\mathrm{Pop}}$ is the **implicit bias** of algorithm $\mathcal{A}$ run on population Pop.

In this problem we consider the implicit bias of the SGD algorithm applied to least squares regression in the case where there are many more parameters than data points. Least squares regression is defined by

$$\Phi[J]^* = \operatorname*{argmin}_{\Phi}\; E_{\langle x,\, y\rangle \sim \mathrm{Train}}\;(\Phi[J]x[J] - y)^2$$

To solve this optimization problem we consider using SGD where Φ is initialized to the zero vector and we then apply the update

$$\begin{aligned}\Phi_{t+1} &= \Phi_t - \eta \nabla_\Phi\, (\Phi^\top x_y - y)^2 \\ &= \Phi_t - 2\eta(\Phi^\top x_t - y)x_t\end{aligned}$$

(a) In the case where $N_{\mathrm{train}} < d$, where d is the dimension of Φ and x, define a linear proper subspace of R^d such that we are guaranteed that Φ_t is in that space for al t.

Solution: Since every update is in the direction of some input vector x_t in the training data, SGD maintains the invariant that Φ_t is some linear combination of the training vectors $x_1, \ldots x_{N_{\mathrm{train}}}$. Since $N_{\mathrm{train}} < d$ the span of the training vectors must be a proper subspace of R^d.

(b) Assume that the training vectors $x_1, \ldots, x_{N_{\mathrm{train}}}$ are linearly independent. In this case it can be shown that there exists a unique solution Φ^* in the space spanned by these vectors for which the square loss of the training data is zero (if these were not independent then we would have more training points than degrees of freedom in the space spanned by the input vectors). Let $b_1, \ldots, b_{N_{\mathrm{train}}}$ be an orthonormal basis for the space spanned by the input vectors. For any $\Phi \in R^d$ define the projection of Φ into the subspace by

$$\begin{aligned}\Phi_\pi &= \sum_i (\Phi^\top b_i) b_i \\ \Phi_\perp &= \Phi - \Phi_\pi\end{aligned}$$

The convergence theorem for SGD now gives that SGD on least squares regression will converge in the limit to Φ^*. Show that SGD applied to least squared regression has a form of implicit bias similar to L_2 regression in that the result Φ^* is the least norm point in R^d for which the square loss of the training data is zero.

Solution: Consider any $\Phi \in R^d$ for which the training loss is zero. The projection Φ_π must also have zero training loss because each training vector can be written as a linear combination of basis vectors and Φ and Φ_π have the same inner product with each basis vector. Therefore $\Phi_\pi = \Phi^*$. Futhermore $\Phi = \Phi_\pi + \Phi_\perp$ and

$$||\Phi||^2 = ||\Phi_\pi||^2 + ||\Phi_\perp||^2 = ||\Phi^*||^2 + ||\Phi_\perp||^2$$

which gives that $||\Phi|| \geq ||\Phi^*||$ as desired.

Problem 6. Generalization Bounds for the realizable case. (25 points) Consider a finite hypothesis class $\mathcal{H}$ and a population distribution Pop on pairs $\langle x, y\rangle$ such that for $\langle x, y\rangle$ drawn from the population and $h \in \mathcal{H}$ we have that h makes a prediction for y which we will write as $h(x)$. The error rate of hypothesis h on the population is defined by

$$\mathrm{Err}_{\mathrm{Pop}}(h) = P_{\langle x,\, y\rangle \sim \mathrm{Pop}}(h(x) \neq y)$$

We draw a training sample Train consisting of N_{Train} pairs $\langle x, y\rangle$ drawn IID from the population.

$$\mathrm{Err}_{\mathrm{train}}(h) = \frac{1}{N_{\mathrm{train}}} \sum_{\langle x,\, y\rangle \in \mathrm{Train}} \mathbf{1}(h(x) \neq y)$$

(a) For a given hypothesis h with error rate ϵ what is the probability that $\mathrm{Err}_{\mathrm{train}}(h) = 0$.

Solution: $(1-\epsilon)^{N_{\mathrm{train}}}$

(b) We now consider a fixed threshold ϵ and consider the hypotheses h satisfying $\mathrm{Err}_{\mathrm{Pop}} \geq \epsilon$. We will call these the "bad" hypotheses.

The simple form of the union bound is

$$P(A \cup B) \leq P(A) + P(B)$$

This can be generalized to

$$P(\exists z\; Q(z)) \leq \sum_z P(Q(z))$$

where $Q(z)$ is any statement about z.

Use your answer to (a) and the union bound to give an upper bound on the probability that there exists a bad hypothesis h with $\mathrm{Err}_{\mathrm{train}}(h) = 0$. You solution should be stated in terms of ϵ, the number of elements $|\mathcal{H}|$ of $\mathcal{H}$, and the number of training pairs N_{train}. Simplify your solution using the inequality $1 - \epsilon \leq e^{-\epsilon}$.

Solution:

$$\begin{aligned}
& P(\exists h \;\; \mathrm{Err}_{\mathrm{Pop}} \geq \epsilon, \; \mathrm{Err}_{\mathrm{train}}(h) = 0) \\
\leq & \sum_{h:\, \mathrm{Err}_{\mathrm{Pop}}(h) \geq \epsilon} P(\mathrm{Err}_{\mathrm{train}}(h) = 0) \\
\leq & |\mathcal{H}|(1-\epsilon)^{N_{\mathrm{train}}} \\
\leq & |\mathcal{H}|e^{-N_{\mathrm{train}}\epsilon}
\end{aligned}$$

(c) Now consider a small positive number δ and solve for ϵ such that the probability that a bad hypothesis has zero training error is less than δ. Your solution gives a value of ϵ such that with probability $1-\delta$ over the draw of the training error all hypothesis with zero training error have population error no lareger than ϵ.

Solution:

$$\begin{aligned}
\delta &= |\mathcal{H}|e^{-N_{\mathrm{train}}\epsilon} \\
\epsilon &= \frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{N_{\mathrm{train}}}
\end{aligned}$$

Problem 7. This problem is on robust loss functions. With a robust loss one identifies "outliers" in the data and "gives up" on modeling the outliers. In particular we can consider the following bounded version of cross-entropy loss

$$\begin{aligned}
\mathcal{L}(\Phi, x, y) &= L_{\mathrm{max}} \; \tanh\left(\frac{-\ln P_\Phi(y|x)}{L_{\mathrm{max}}}\right) \\
\tanh(z) &= \frac{2}{1+e^{-2z}} - 1.
\end{aligned}$$

For $z \geq 0$ we have $\tanh(z) \geq 0$ and we have that the above robust loss is non-negative and can never be larger than L_{max}.

(a) Consider the function $L_{\mathrm{max}} \; \tanh(\frac{z}{L_{\mathrm{max}}})$. Use a first order Taylor expansion of the tanh function about zero to show that for $|z| << L_{\mathrm{max}}$ we have

$$L_{\mathrm{max}} \; \tanh\left(\frac{z}{L_{\mathrm{max}}}\right) \approx z$$

This implies that the robust cross entropy loss is essentially equal to the cross entropy loss when the cross entropy loss is small compared to $L_{\max}$.

Solution: The first order Taylor expansion of the tanh function about zero is

$$\tanh(u) \approx u$$

yielding the desired result.

(b) Consider the case where the cross-entropy loss is large compared to $L_{\max}$. For $z >> 1$ we have that the derivative $\tanh'(z)$ is essentially zero. What parameter update is made on a training point whose cross entropy loss is large compared to $L_{\max}$ if we model $\tanh'(z) = 0$ in such cases.

Solution: The update on a data point (x, y) is

$$\Phi^{t+1} = \Phi^t - \eta \nabla_\Phi \mathcal{L}(\Phi, x, y)$$

At a point where the derivative of the sigmoid is essentially zero this update will be essentially zero. So "outliers" do not effect the model parameters.

(c) Look up the PAC-Bayesian generalization guarantee that is stated in terms of the L_2 norm of the weight vector. Explain why the robust loss function comes with a better PAC-Bayesian generalization guarantee. Intuitively, the improvement in generalization is due to insensitivity to "outliers" (or things the model cannot understand).

Solution: The L_2 PAC-Bayeisan guarantee in the notes is

$$\mathcal{L}_\sigma(\Phi) \leq \frac{10}{9}\left(\hat{\mathcal{L}}_\sigma(\Phi) + \frac{5L_{\max}}{N}\left(\frac{||\Phi - \Phi_{\mathrm{init}}||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right)$$

Reducing $L_{\max}$ both reduces $\hat{L}_\sigma(\Phi)$ and reduces the penalty for the model complexity (the norm squared of the distance from the initialization).

(d) Curriculum learning is the idea that one first learns how to solve easy problems and then gradually learns ever harder problems. At a high informal level describe a learning algorithm based on the above robust loss function which can be intuitively motivated as curriculum learning.

Solution: Easier problems should correspond to cases where the cross entropy loss can be made small. So setting $L_{\max}$ to be smallish will focus the model on the easy problems while ignoring the hard problems. Gradually increasing $L_{\max}$ will gradually pay more attention to the harder problems.

Another possible answer is that holding $L_{\max}$ fixed will focus first on the easy problems ignoring the hard problems but as the understanding of easy problems improves the harder problems become easy problems and we automatically gradually pay attention to harder and harder problems even with a fixed value of $L_{\max}$.

Problem 8. This problem is on PAC-Bayes bounds for classifiers built on CLIP using **prompt engineering**. CLIP is a joint probability model on images and English descriptions (image captions). Clip is trained on a large corpus of captioned images drawn from the web and defines a probability distribution over captions c given an image x. We can use CLIP for image classification (as in ImageNet) using "prompt engineering". A "prompt" is caption specific to an image label. For example the caption "this is an image of a cat" for the label "cat" or "this is an image of a dog" for the label "dog". For each image class y we have a prompt (hypothetical caption) $c(y)$. We can then label an image x with class $\hat{y}$ using the rule

$$\hat{y}(x) = \underset{y}{\mathrm{argmax}}\; P_{\mathrm{CLIP}}(c(y)|x)$$

Suppose that we search (somehow) over the captions $c(y_1), \ldots, c(y_n)$ assigned to the n image classes $y_1, \ldots, y_n$ to find a set of captions minimizing the error rate (0-1 loss) on a set of N labeled training images. Let $\hat{\mathcal{L}}$ be the error rate on the training data. Also suppose that CLIP assigns a prior probability $P_{\mathrm{CLIP}}(c)$ to any caption c independent of any image. Consider the PAC-Bayes bound on generalization loss for predictive rule h where the bound is guaranteed to hold for all h with probability at least $1 - \delta$.

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N_{\mathrm{Train}}}\left(-\ln P(h) + \ln\frac{1}{\delta}\right)\right)$$

Apply this rule to the CLIP image classifier using CLIP's "prior probability" on the caption space.

Solution:

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5}{N}\left(\left(\sum_y -\ln P_{\mathrm{CLIP}}(c(y))\right) + \ln\frac{1}{\delta}\right)\right)$$

I am not proposing that searching over all captions is a good idea. Some narrower prior is called for.

Problem 1: Generalization Bounds and The Lottery Ticket Hypothesis. Suppose that we want to construct a linear classifier (a linear threshold unit) for binary classification defined by

$$\hat{y}_\alpha(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^d \alpha_i f_i(x) >= 0 \\ \\ -1 & \text{otherwise} \end{cases}$$

where each α_i is a scalar weight, $f_i(x)$ is a scalar value, and the functions f_i are (random) features constructed independent of any observed values of x or y. We will assume a population distribution Pop of pairs $\langle x, y\rangle$ with $y \in \{-1, 1\}$ and a training set Train of N pairs drawn IID from pop.

We can define both test and train losses (error rates).

$$\hat{\mathcal{L}}(\alpha) \;=\; E_{x,y\sim\text{Train}}\;\mathbf{1}[\hat{y}_\alpha(x_i) \neq y_i]$$

$$\mathcal{L}(\alpha) \;=\; E_{x,y\sim\text{Pop}}\;\;\mathbf{1}[\hat{y}_\alpha(x_i) \neq y_i]$$

Assume finite precision arithmetic so that we have discrete rather than continuous possible values of α. The course slides state that for any (prior) distribution P on the values of α we have that with probability at least $1-\delta$ over the draw of the training data the following holds simultneously for all α.

$$\mathcal{L}(\alpha) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(\alpha) + \frac{5L_{\max}}{N}\left(-\ln P(\alpha) + \ln\frac{1}{\delta}\right)\right)$$

We will now incorporate the lottery ticket hypothesis into the prior distribution on α by assuming that low training error can be achieved with some small subset of the random features. More formally, we define a prior favoring sparse α — cases where most weights are zero.

(a) To define $P(\alpha)$, first define a prior probability distribution $P(s)$ over the number s of nonzero values.

Solution: There are of course many solutions. A uniform distribution on the numbers from 1 to d will work giving $P(s) = 1/d$. Another possibility is $P(s) = \epsilon(1-\epsilon)^s$ which defines a distribution on all $s \geq 0$.

(b) Given a specified number s of nonzero values, define a probability distribution $P(U|s)$ where U is a subset of the random features with $|U| = s$.

Solution: A reasonable choice here is a uniform distribution on the $\binom{d}{s}$ possibilities giving $P(U|s) = 1/\binom{d}{s}$.

(c) Assuming that each nonzero value is represented by b bits, give a probability distribution over $P(\alpha|U,s)$.

Solution: Here we can use the uniform distribution on the 2^{bs} ways of assigning numbers to the s nonzero weights in α giving $P(\alpha|U,s) = P(\alpha|U) = 2^{-bs}$.

(d) Combine (a), (b) and (c) to define $P(\alpha)$.

Solution: Under $P(s) = 1/d$ we get $P(\alpha) = \frac{1}{d\binom{d}{s}2^{bs}}$ and using $\binom{d}{s} \leq d^s$ we get $P(\alpha) \geq \frac{1}{dd^s2^{bs}} = \frac{1}{d^{s+1}2^{bs}}$.

Under $P(s) = \epsilon(1-\epsilon)^s$ we get $P(\alpha) = \frac{\epsilon(1-\epsilon)^s}{\binom{d}{s}2^{bs}}$ and using $\binom{d}{s} \leq d^s$ $P(\alpha) \geq \frac{\epsilon(1-\epsilon)^s}{d^s2^{bs}}$

(e) Plug your answer to (c) into the above generalization bound to get a bound in terms of the number of random features d, the number s of nonzero values of α, and the number b of bits used to represent each nonzero value and any additional parameters used in defining your distributions.

Solution: Under $P(s) = 1/d$ we get

$$\begin{aligned}
\mathcal{L} &\leq \frac{10}{9}\left(\hat{\mathcal{L}} + \frac{5}{N}\left(\ln d + \ln \binom{d}{s} + sb\ln 2 + \ln\frac{1}{\delta}\right)\right) \\
&\leq \frac{10}{9}\left(\hat{\mathcal{L}} + \frac{5}{N}\left((s+1)\ln d + sb\ln 2 + \ln\frac{1}{\delta}\right)\right)
\end{aligned}$$

Under $P(s) = \epsilon(1-\epsilon)^s$ we get

$$\begin{aligned}
\mathcal{L} &\leq \frac{10}{9}\left(\hat{\mathcal{L}} + \frac{5}{N}\left(\ln\frac{1}{\epsilon} + s\ln\frac{1}{1-\epsilon} + \ln \binom{d}{s} + sb\ln 2 + \ln\frac{1}{\delta}\right)\right) \\
&\leq \frac{10}{9}\left(\hat{\mathcal{L}} + \frac{5}{N}\left(\ln\frac{1}{\epsilon} + s\ln\frac{1}{1-\epsilon} + s\ln d + sb\ln 2 + \ln\frac{1}{\delta}\right)\right)
\end{aligned}$$

Note that in either case the bound is logarithmic in d allowing d to be extremely large. The choice of the uniform distribution for s is simpler and gives a completely satisfactory result. However there are regimes in which the second prior on s is very slightly better.