A SURVEY OF DEEP LEARNING FOR MATHEMATICIANS

TONY FENG

3. Statistical inference

3.1. **Central Limit Theorem.** Let $X_1, X_2, \ldots$ be i.i.d. random variables with (finite) mean μ.

Slogan 3.1.1. The *Law of Large Numbers* says that the random variables

$$\overline{X}_n := \frac{X_1 + \ldots + X_n}{n}$$

converge to the mean μ as $n \to \infty$.

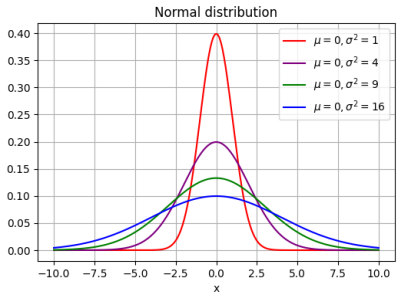
One precise formulation of Slogan 3.1.1 is that for any fixed $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0.$$

(This is called the "Weak Law of Large numbers"; it gives a type of convergence called *convergence in probability.*)

Given the Law of Large Numbers, it is natural to ask about the random fluctuation of $\overline{X}_n$ around its mean μ. Suppose X_i has variance σ^2. Recall that the *Gaussian* (or *normal*) distribution with mean μ and variance σ^2, denoted $\mathcal{N}(\mu, \sigma^2)$, has density function

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Slogan 3.1.2. Suppose the X_i have finite mean σ^2. The *Central Limit Theorem* says that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_n - \mu).$$

converges in distribution to a Gaussian distribution with mean 0 and variance σ^2.

The Central Limit Theorem explains the "universality" of the Gaussian distribution: it is the limiting behavior of the average of independent (identically distributed) random variables. Note that $\mathrm{Var}(\sum_{i=1}^{n}(X_n - \mu)) = \sum_{i=1}^{n} \mathrm{Var}(X_i - \mu) = n\sigma^2$, hence why $\sqrt{n}$ is the correct normalization factor to see an interesting result. This is responsible for the famous

Date: Sep 8, 2025.

"square root cancellation" principle, asserting that an accumulation of n random fluctuations in $[-1,1]$ can be expected to reach the order of magnitude of $\sqrt{n}$.

We will give a partial derivation of the Central Limit Theorem (one that is valid under additional technical assumptions). Recall that the *moment generating function* of a random variable X is the formal series

$$M_X(t) = \mathbb{E}[e^{tX}] = 1 + t\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2] + \ldots + \frac{t^k}{k!}\mathbb{E}[X^k] + \ldots$$

Example 3.1.3. The moment generating function of $X \sim \mathcal{N}(0,\sigma^2)$ is

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2\sigma^2}x^2}\, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x^2 - 2t\sigma^2 x)}\, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}((x-t\sigma^2)^2 - t^2\sigma^4)}\, dx \\ &= e^{\frac{t^2\sigma^2}{2}}. \end{aligned}$$

More generally, this implies by a translation argument that the moment generating function of $X \sim \mathcal{N}(\mu,\sigma^2)$ is

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

This is essentially a *characterization* of the Gaussian distribution: the logarithm of its moment generating function is a quadratic polynomial in t, of the form $\mu t + \frac{\sigma^2 t^2}{2}$.

Now let's derive the Central Limit Theorem under the assumption that $X_1, X_2, \ldots$ are i.i.d. random variables with well-defined moment generating function[1] $M_{X_i}(t)$. Without loss of generality, assume that $\mu = 0$, so that

$$M_{X_i}(t) = 1 + \frac{\sigma^2 t^2}{2} + O(t^3)$$

We have

$$M_{\frac{1}{\sqrt{n}}\sum X_i}(t) = \mathbb{E}\left[e^{t\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i}\right] = \mathbb{E}[e^{tX_i/\sqrt{n}}]^n = M_{X_i}(t/\sqrt{n})^n.$$

Taking logarithms, we then see that

$$\ln M_{\overline{X}_n}(t) = n \ln M_{X_i}(t/\sqrt{n}) = n \ln\left(1 + \frac{\sigma^2 (t/\sqrt{n})^2}{2} + O((t/\sqrt{n})^3)\right)$$

Using the Taylor series of $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots$, we find that all but the lowest order term vanish as $n \to \infty$, so that

$$\lim_{n\to\infty} \ln M_{\overline{X}_n}(t) = \frac{\sigma^2 t^2}{2},$$

indicating a Gaussian distribution. □

[1] This does not always exist. In general, we should work instead with the *characteristic function* $\mathbb{E}[e^{itX}]$, which always exists.

3.2. **Estimators.** Let P be a probability distribution.

Definition 3.2.1. A *random sample* from the distribution $f(x)$ is a sequence of random variables $X_1, \ldots, X_n$ which are i.i.d. with distribution P.

Definition 3.2.2. A *statistic* of a random sample $(X_1, \ldots, X_n)$ is a random variable of the form $W = T(X_1, \ldots, X_n)$ for some function T.

Example 3.2.3 (Sample mean)**.** The *sample mean* of $X_1, \ldots, X_n$ is the statistic

$$\overline{X} := \frac{1}{n}(X_1 + \ldots X_n).$$

Note the difference between the *mean* of the X_i, which is a number, and the sample mean, which is itself a random variable.

In statistical inference, we will find ourselves in the following kind of setup. Suppose we have family of probability distributions P_θ, depending on a parameter θ. An *estimator* of a random sample $(X_1, \ldots, X_n)$ is a function $W(X_1, \ldots, X_n)$, which is intended to infer the parameter θ. Mathematically, an estimator is identical to a statistic; the difference between the two terms lies in their psychological connotations: we think of an estimator as "trying" to estimate something, whereas a statistic does not necessarily have any attached "purpose".

Definition 3.2.4. Below, we denote by $\mathbb{E}_\theta[W]$ the expected value of W according to the distribution P_θ. The *bias* of an estimator W of a parameter θ is

$$\text{Bias}_\theta(W) := \mathbb{E}_\theta[W] - \theta.$$

An estimator whose bias is 0 is called *unbiased*.

Example 3.2.5 (Estimating the mean)**.** By linearity of expectation,

$$\mathbb{E}[\overline{X}] = \frac{1}{n}(\mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]) = \mathbb{E}[X_i]$$

since the X_i are i.i.d. In other words, the sample mean is an unbiased estimator of the mean.

Example 3.2.6. Let $X_1, \ldots, X_n$ be a random sample. For each $i = 1, \ldots, n$, we can take X_i as an estimator of the mean of the X_i. This is obviously an unbiased estimator.

Example 3.2.7 (Estimating the variance)**.** Recall that the *variance* of X_i is

$$\mathbb{E}[(X_i - \mathbb{E}_\theta[X_i])^2] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2.$$

Let's define the (ad hoc terminology) "naive sample variance" to be

$$S^2_{\text{naive}}(X_1, \ldots, X_n) := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Is this an unbiased estimator of the variance? Let's calculate $\mathbb{E}[S^2_{\text{naive}}(X_1, \ldots, X_n)]$. First, some algebra gives

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_i \Big((X_i - \mu) - (\overline{X} - \mu)\Big)^2 \\ &= \sum_i (X_i - \mu)^2 - 2\sum_i (X_i - \mu)(\overline{X} - \mu) + \sum (\overline{X} - \mu)^2.\end{aligned}$$

Using that $\sum_i (X_i - \mu) = n(\overline{X} - \mu)$, this expression simplifies as

$$\sum_i (X_i - \mu)^2 - n(\overline{X} - \mu)^2.$$

Then taking expectation, we find that

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=1}^n (X_i - \overline{X})^2\Big] &= \mathbb{E}\Big[\sum_i (X_i - \mu)^2 - n(\overline{X} - \mu)^2\Big] \\ &= n\operatorname{Var}[X_i] - n\frac{\operatorname{Var}[X_i]}{n} = (n-1)\operatorname{Var}[X_i].\end{aligned}$$

Dividing by n, we obtain

$$\mathbb{E}[S^2_{\text{naive}}(X_1, \dots, X_n)] = \frac{n-1}{n}\operatorname{Var}[X_i].$$

In particular, the naive sample variance is *not* unbiased (unless $\operatorname{Var}[X_i]$ happens to be 0, in which case there is no randomness anyway). Furthermore, we see what is needed to correct it: define the *sample variance*

$$S^2(X_1, \dots, X_n) := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{n}{n-1} S^2_{\text{naive}}.$$

Then the same calculation shows that $S^2(X_1, \dots, X_n)$ is an unbiased estimator of the variance.

What happened here? Qualitatively, we saw that the naive sample variance is an *underestimate* of the variance. For example, looking at the case $n = 1$, we see that the naive sample variance would always be 0, which is clearly an underestimate, while the sample variance is undefined (although this may appear mathematically disturbing at first, it is actually reasonable: a single sample is not enough to conclude anything about variance).

3.3. **Maximum Likelihood Estimation.** Suppose we have family of probability distribution, parametrized by θ. We denote the PMF/PDF by $p(x|\theta)$.

- The probability of an outcome x given parameter θ is $p(x|\theta)$.
- The *likelihood* of a parameter θ given an outcome x is $L(\theta|x) := p(x|\theta)$.

Note that the right side "$p(x|\theta)$" of these formulas looks the same; the difference lies in what is being held fixed and what is viewed as variable.

Slogan 3.3.1. In practical terms, we use probability to predict outcomes from a given model, while we use likelihood to infer a model from a given outcome.

Definition 3.3.2. Suppose we are given an outcome x and want to infer θ. The *maximum likelihood estimate* of θ is

$$\hat{\theta} := \arg\max_\theta L(\theta|x) = \arg\max_\theta p(x|\theta).$$

Example 3.3.3. Let $X_1, X_2, \dots, X_n$ be a random sample from a Bernoulli distribution Bernoulli(θ). Then for $\underline{x} = (x_1, \dots, x_n) \in \{0,1\}^n$ we have

$$L(\theta|\underline{x}) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^k(1-\theta)^{n-k} \quad \text{where } k = \#\{i\colon x_i = 1\}.$$

We could differentiate this directly, but we will instead leverage the key observation that because of monotonicity of log, *maximizing the likelihood is equivalent to maximizing its logarithm*. Now, the "log-likelihood" $\ln L(\theta|\underline{x})$ has a more convenient expression,

$$\ln L(\theta|\underline{x}) = k\ln(\theta) + (n-k)\ln(1-\theta). \tag{3.3.0.1}$$

Now we can easily differentiate to solve for the unique critical point $\hat{\theta} = k/n$. A little more work shows that this is the global maximum of (3.3.0.1), so we conclude that the maximum likelihood estimate is (the obvious guess) $\hat{\theta} = k/n$.

Example 3.3.4. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Then for $\underline{x} = (x_1, \ldots, x_n)$, we have

$$L(\mu, \sigma^2|\underline{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2}.$$

To maximize $L(\mu, \sigma^2|\underline{x})$, we will again use the trick of looking instead at the log-likelihood,

$$\log L(\mu, \sigma^2|\underline{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Let's examine the critical points. Differentiating with respect to the parameters, we find that

$$\frac{d}{d\mu}\ln L(\mu, \sigma^2|\underline{x}) = \frac{1}{\sigma^2}\left(\sum_i (x_i - \mu)\right), \tag{3.3.0.2}$$

and

$$\frac{d}{d\sigma^2}\ln L(\mu, \sigma^2|\underline{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{3.3.0.3}$$

Setting (3.3.0.2) equal to 0 gives $\boxed{\hat{\mu} = \frac{1}{n}\sum_i x_i,}$ and then setting (3.3.0.3) equal to 0 gives $\boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.}$ Note that these are what we called the *sample mean* and *naive sample variance*, respectively. With a bit more work (to verify that these $\hat{\mu}$ and $\hat{\sigma}^2$ give global maxima), we find that this provides the maximum likelihood estimates. In particular, the MLE can be biased, as we see in this case by Example 3.2.7.

Remark 3.3.5. One *weakness* of MLE is that it can be highly unstable as a function of the data, which in practice is noisy. The following example (due to Olkin et al.) is [**?**, p.297, Example 7.2.2]. Suppose we sample from a Binomial(k, p) distribution with unknown k and p. Consider the two datasets

- $[16, 18, 22, 25, 27]$
- $[16, 18, 22, 25, 28]$

For the first, the MLE of k is $\hat{k} = 99$, while for the second the MLE of k is $\hat{k} = 190$.

3.3.1. *Relation to cross-entropy.* Let p and q be two probability distributions on the same sample space (which we conflate with their PMF/PDF). Then the *cross-entropy* of p, q is

$$H(p, q) := \mathbb{E}_p[\log(1/q)] = \mathbb{E}_p[-\log q].$$

Example 3.3.6. Supposing that p and q are discrete for simplicity, this unpacks to

$$H(p,q) = -\sum_i p_i \log(q_i).$$

As with the KL divergence, there are some edge cases in interpreting this formula (Remark **??**): if $p_i = 0$, then the summand is omitted, while if $p_i \neq 0$ but $q_i = 0$, then $H(p,q) = \infty$.

In the case of discrete random variables, it is clear that $H(p,q) \geq 0$, since each $-p_i \log(q_i) \geq 0$. This tells us that we can reasonably think of $H(p,q)$ as being some measure of "distance" between p and q. (Note however that $H(p,q) \neq H(q,p)$. We'll come back to this later.) But what is it trying to capture?

Let's place ourselves in the mindset of logistic regression. Think of p as being the empirical distribution given by data (corresponding to x), and q as being the model parameter (corresponding to θ). Then the Maximum Likelihood Estimate of the parameter maximizes the likelihood of the observed data under the model q.

Imagine for concreteness that the empirical distribution p arises from N total outcomes, so outcome ω_i comes up Np_i times (if it helps psychologically, you can imagine that N is divisible enough so that Np_i is an integer). Under our model q, the probability of this happening is $q_i^{Np_i}$, so the maximum likelihood estimate is

$$\hat{q} = \arg\max_q \Big(\prod_i q_i^{Np_i} \Big).$$

Again, it is equivalent to maximize the log-likelihood,

$$\hat{q} = \arg\max_q \Big(\prod_i q_i^{Np_i} \Big) = \arg\max_q \log \Big(\prod_i q_i^{Np_i} \Big) = \arg\max_q \Big(\sum_i Np_i \log q_i \Big).$$

Now we see that the N is irrelevant, so we can pull it out. Also, to put this in our standard "loss minimization" framework, we can trade the $\arg\max$ for $\arg\min$ if we introduce a negative sign. This means that the maximum likelihood estimate can be written equivalently

$$\hat{q} = \arg\min_q H(p,q)$$

and this suggests the cross-entropy $H(p,q)$ as our loss function. In other words, *minimizing cross-entropy loss is solving for the maximum likelihood estimate!*

Remark 3.3.7. From this analysis, we see that $H(p,q)$ reflects "the probability of empirical data p under the model q". This description is not symmetric under exchanging $p \leftrightarrow q$. Inspecting the formula $-\sum p_i \log q_i$, we see that summands where q_i is small but p_i is non-negligible contribute a large value. This guides us to the heuristic principle:

Slogan 3.3.8. Cross-entropy loss punishes predictions that are "confident but wrong" that an outcome will not occur.

Example 3.3.9. In classification problems, it is common to train a neural net that computes a probability distribution q over possible classes, and then outputs the most probable class. Suppose the data is of the form $\{(x_i, y_i)\}$ where y_i is the ground truth label on the input x_i. The loss function for a datapoint (x_i, y_i) would then by cross-entropy $H(p,q)$ where p is the delta function on the true class y_i, and q is the model's probability distribution. This is simply

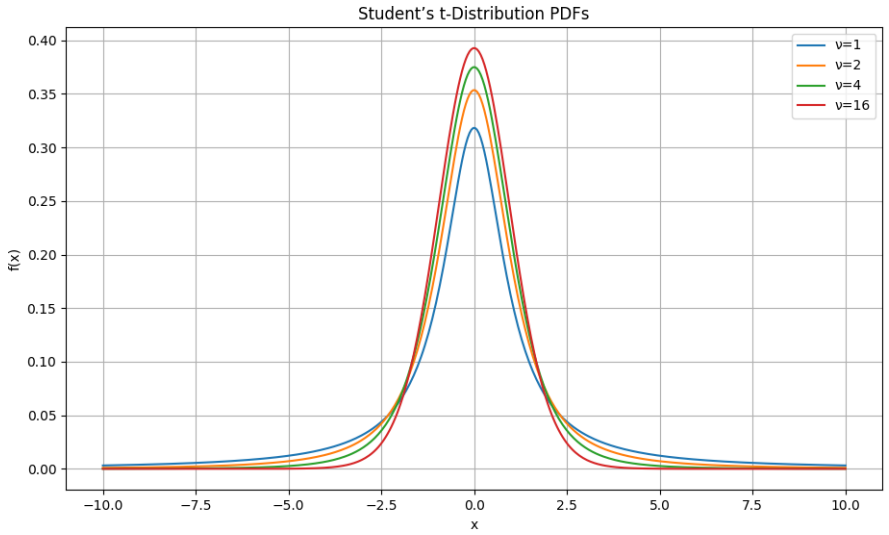
$$H(p,q) = -\log q(y_i).$$

By contrast, $H(q,p)$ would be infinite as long as q is supported on some outcome other than y_i, which would typically be the case. Therefore, $H(q,p)$ would not be an appropriate loss function.

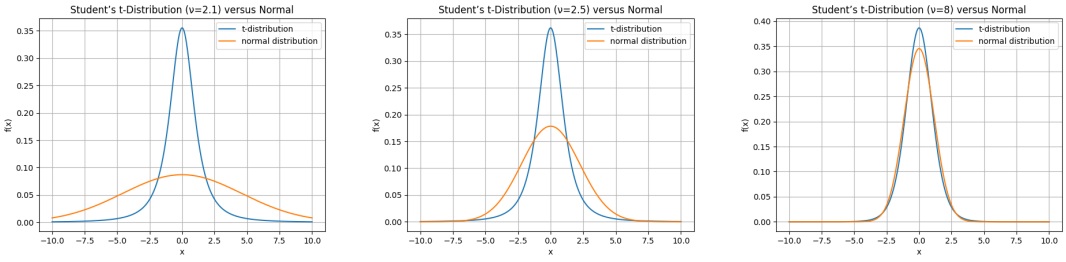
Example 3.3.10. Student's[2] "t_ν-distribution" depends on a parameter ν, and has PDF

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

The graph of f_ν is "bell-shaped", roughly similar in appearance to a normal distribution.



For $\nu > 2$, f_ν has mean 0 and finite variance $\frac{\nu}{\nu-2}$. Let's compare to the normal distributions with these means and variances.



What do the cross-entropies look like in each of these situations? The following table is the result of numerical (truncated) integration over $x \in [-500, 500]$.

$H(t_{2.1}, \mathcal{N}(0,21))$	2.5738
$H(t_{2.5}, \mathcal{N}(0,5))$	2.0782
$H(t_8, \mathcal{N}(0,4/3))$	1.5625

$H(\mathcal{N}(0,21), t_{2.1})$	3.726
$H(\mathcal{N}(0,5), t_{2.5})$	2.4343
$H(\mathcal{N}(0,4/3), t_8)$	1.5722

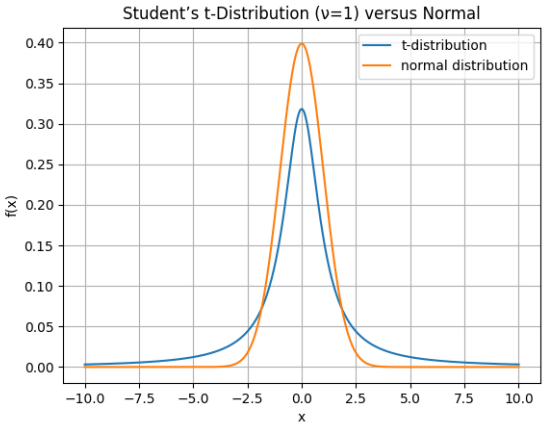
[2]Student is a name, which a pseudonym used by William Sealy Gosset.

For $\nu = 2.1, 2.5$ we see that $H(t_\nu, \mathcal{N}(0,\sigma^2))$ is noticeably smaller than $H(\mathcal{N}(0,\sigma^2), t_\nu)$, while they are pretty close for $\nu = 8$. Let's try to rationalize how we could have guessed this from looking at the graphs. Intuitively, taking the t_ν-distribution as the model and the normal distribution as the data, we might imagine that it gets penalized heavily in the low-probability tails where the normal distribution lies above it. With the roles switched, the region where the t_ν-distribution lies significantly above the normal is also the highest probability region for the normal distribution.

That being said, human intuition can misleading – especially at extreme scale or high-dimensional settings – and there is a delicate balance between casting things in terms of intuition and just trusting the mathematics. Indeed, from looking at the PDFs it is clear that Student's t-distribution is actually *heavier* in the tails asymptotically than the normal distribution, since it decays as a power of x, whereas the density function of the normal distribution decays exponentially. More precisely, $f_\nu(x)$ decays like $x^{-(\nu+1)}$ as $|x| \to \infty$, whereas

$$-\log f_{\mathcal{N}(\mu,\sigma^2)}(x) \asymp x^2.$$

In particular, we can see that for $\nu \leq 2$, the cross-entropy $H(t_\nu, \mathcal{N}(\mu,\sigma^2))$ will *diverge*. Let's examine $\nu = 1$ more closely; in this case, the variance of t_ν diverges because the PDF decays like x^{-2}, but we can still compare it to the standard normal distribution.



It's not obvious from the picture, but the t-distribution has heavier tails than the normal distribution, leading to $H(t_1, \mathcal{N}(0,1)) = \infty$, while $H(\mathcal{N}(0,1), t_1) \approx 1.6782$. Intuitively, this is because the cross-entropy penalizes the tail events of the t-distribution, which the normal distribution "thinks" should almost never occur.

3.3.2. *Relation to KL divergence.* Let p, q be two probability distributions on the same outcome space. Recall that the KL divergence of p, q is

$$\mathrm{KL}(p,q) := \mathbb{E}_p[\log(p/q)] = H(p,q) - H(p).$$

One sees both the KL divergence and cross-entropy used in machine learning, often interchangeably. What is the difference between them? Recall that in logistic regression, p is the empirical distribution determined from data, and q is the model. Hence $\mathrm{KL}(p,q)$ differs from $H(p,q)$ by the constant $H(p) = -\sum p_i \log p_i$, which is independent of the model. Since the addition of this constant doesn't affect the gradient of the loss function, these two loss functions are completely equivalent for the purpose of backpropagation.

3.4. **Bayesian estimation.** Thus far we have discussed what are called "point" estimates: a numerical estimate for a parameter θ. According to the Bayesian perspective of statistics, one should really estimate a "posterior *distribution*" on θ by using the data $\mathcal{D}$ to calculate an update to the "prior distribution". Given a "prior distribution" $P(\theta)$, this can be done using Bayes' rule:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}.$$

Example 3.4.1. Suppose the Warriors play the Lakers 3 times in the NBA finals, and win 2 of them. If they play again, what is the probability that the Warriors will win?

This question does not have a single "correct" answer; we just have to propose a model and see how it performs. Let's model the situation by assuming that the outcome of each game is i.i.d. and that the Warriors have win probability $\theta \in [0, 1]$; in other words, the outcomes of games independent samples from a Bernoulli distribution. In Example 3.3.3 we saw that the MLE of θ is then $\hat{\theta} = 2/3$.

Note that if the Warriors had won all 3 past games, the MLE would be $\hat{\theta} = 1$, or 100% chancing of winning Game 4. This may make you uncomfortable, indicating that you have *prior beliefs* about θ which are not reflected in this model.

Let's consider instead a Bayesian perspective. We have to start by choosing a prior distribution on θ, and a the most natural one to take seems to be the uniform distribution. Let $\mathcal{D}$ be the data that the Warriors won 2 of the first 3 games, and A the event that they win game 4. Then the prior distribution of $P(A) = \theta$ is uniform, so that its density function $p(\theta)$ is the indicator function of $[0, 1]$. By Bayes' rule, the *posterior distribution* $P(A|\mathcal{D})$ has density function

$$p(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)p(\theta)}{P(\mathcal{D})}.$$

We have $P(\mathcal{D}|\theta) = \binom{3}{2}\theta^2(1-\theta)$, and then

$$P(\mathcal{D}) = \int_0^1 P(\mathcal{D}|\theta)d\theta = \binom{3}{2}\int_0^1 \theta^2(1-\theta)\,d\theta. \tag{3.4.0.1}$$

At this point, it is useful to remember that the *Beta distribution* $B(\alpha, \beta)$ is the probability distribution with density function proportional to $x^{\alpha-1}(1-x)^{\beta-1}$, and normalization constant determined by the formula[3]

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

From this, we see that (3.4.0.1) comes out to $\binom{3}{2}\frac{2!1!}{4!} = \frac{1}{4}$. Thus

$$p(\theta|\mathcal{D}) = \begin{cases} 12\theta^2(1-\theta) & \theta \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

Finally, we can use this to calculate the probability that the Warriors win the next game under our model: it is

$$P(A|\mathcal{D}) = \int_0^1 P(A|\theta)p(\theta|\mathcal{D})\,d\theta = \int_0^1 \theta f(\theta|\mathcal{D})\,d\theta = 12\int_0^1 \theta^3(1-\theta) = 12\frac{3!}{5!} = \frac{3}{5}.$$

[3]If α and β are not integers, then the interpretation of this formula is "$(\alpha-1)!$" $= \Gamma(\alpha)$.

Remark 3.4.2. The θ that maximizing $P(\theta|\mathcal{D})$ is called the *maximum a posteriori* (MAP) estimate. In this example, the MAP estimate (with uniform prior) happens to agree with the MLE estimate $2/3$, but it could have differed if the prior distribution was non-uniform.

3.4.1. *Relation to loss functions.* Suppose we have data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ which we are trying to model with a function $y = f_\theta(x)$. Consider the L^2-regularized loss function

$$L(\theta) = \sum_{i=1}^n (y_i - f_\theta(x_i))^2 + ||\theta||_2^2.$$

This can be interpreted in terms of MAP estimation with Gaussian prior and noise. Indeed, Bayes' rule says that the posterior density $p(\theta|\mathcal{D})$ is given by

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}.$$

Since $p(\mathcal{D})$ is a constant independent of θ, we can ignore it for the purposes of optimizing θ. Hence it is equivalent to find the $\hat{\theta}$ which minimizes

$$-\ln p(\theta|\mathcal{D}) = -\ln p(\mathcal{D}|\theta) + \ln p(\theta).$$

If we choose the prior distribution on $p(\theta)$ to be the product of independent mean-zero Gaussian distributions, then we have

$$-\ln p(\theta) = \sum_i \left(\frac{\theta_i^2}{2\sigma^2}\right)$$

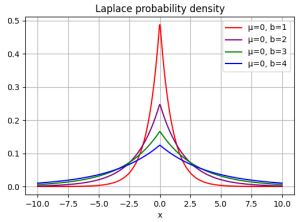
which reproduces the regularization term $||\theta||_2^2$. If we assume that the data $\{(x_i, y_i)\}_{i=1}^n$ is of the form $y_i = f_\theta(x_i) + \epsilon_i$ where the error terms ϵ_i are drawn i.i.d. from a Gaussian of mean 0, then the other term $-\ln p(\mathcal{D}|\theta)$ agrees (up to constants) with the MSE loss

$$\sum_i (y_i - f_\theta(x_i))^2.$$

Slogan 3.4.3. The L^2-regularized loss function solves for the MAP estimate for data which is drawn from a distribution with Gaussian errors, and with Gaussian prior on the parameters. (These are natural distributions to assume, by the universality of the Gaussian.)

Example 3.4.4. The *Laplace distribution* with mean μ and variance $2b^2$ has probability density function

$$f(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$



By the same discussion, we see that L^1-regularization can be interpreted in terms of MAP estimation with a Laplacian prior distribution on weights.

3.5. **Evaluating estimators.** Let W be an estimator for a parameter θ.

- The *mean absolute error* (MAE) of W is $\mathbb{E}_\theta[|W-\theta|]$, viewed as a function of θ.
- The *mean squared error* (MSE) of W is $\mathbb{E}_\theta[(W-\theta)^2]$, viewed as a function of θ.

The MSE is somewhat preferred, as it is analytically convenient ($(W-\theta)^2$ is better to differentiate as a function of θ). It also has the benefit of being "interpretable" according to the decomposition

$$\begin{aligned}\mathbb{E}_\theta[(W-\theta)^2] &= \mathbb{E}_\theta[(W-\mathbb{E}_\theta[W]+\mathbb{E}_\theta[W]-\theta)]^2 = \mathbb{E}_\theta[(W-\mathbb{E}_\theta[W])^2] + (\mathbb{E}_\theta[W]-\theta)^2 \\ &= \mathrm{Var}_\theta(W) + \mathrm{Bias}_\theta(W)^2\end{aligned}$$

where we recall the notion of bias from Definition 3.2.4. The variance can be viewed as a measure of the precision versus noisiness of the estimator, while the bias can be viewed as a measure of accuracy.

Example 3.5.1. Let $X_1,\dots,X_n$ be a random sample from a distribution with mean μ and variance σ^2. We saw in Example 3.2.5 that the sample mean $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator of the mean μ of the X_i. Therefore, the MSE of this estimator is $\boxed{\mathrm{Var}[\overline{X}] = \frac{\sigma^2}{n}.}$

Example 3.5.2. Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. We can simply take X_1 as an estimator for μ. Tautologically, this is unbiased, and tautologically its MSE is $\mathbb{E}[(X_1-\mu)^2] = \mathrm{Var}[X_1] = \sigma^2$. Compared to the sample mean, the variance is much higher, making it a "worse" estimator than the sample mean. However, it has a benefit: it is faster to compute, as we just need to draw a single sample. Therefore, we will use this estimator when training certain computationally intensive models, such as Diffusion Models, at scale.

Example 3.5.3. We move on to consider estimators of the variance. We saw in Example 3.2.5 that the sample variance

$$S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2$$

is an unbiased estimator of the variance. Therefore, $\mathrm{MSE}(S^2) = \mathrm{Var}_{\sigma^2}(S^2)$. Through some tedious algebra, one can calculate this to be

$$\mathrm{MSE}(S^2) = \mathbb{E}[(S^2-\sigma^2)^2] = \frac{2\sigma^4}{n-1}.$$

Now suppose that the X_i are Gaussian. Then in Example 3.3.4 we saw that the MLE estimate is what we called in Example 3.2.5 the *naive sample variance*

$$S^2_{\text{naive}} = \frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2.$$

In particular, the MLE estimate is biased, with

$$\mathrm{Bias}(S^2_{\text{naive}})^2 = (\mathbb{E}[S^2_{\text{naive}}]-\sigma^2)^2 = \frac{\sigma^4}{n^2}.$$

As for the variance, we have

$$\mathrm{Var}(S^2_{\text{naive}}) = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}(S^2) = \frac{2(n-1)\sigma^4}{n^2}.$$

Hence the MSE of S^2_{naive} is

$$\text{MSE}(S^2_{\text{naive}}) = \frac{\sigma^4}{n^2} + \frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.$$

To summarize, we have seen that

$$\text{MSE}(S^2_{\text{naive}}) = \frac{(2n-1)\sigma^4}{n^2} < \frac{2\sigma^4}{n-1} = \text{MSE}(S^2).$$

Thus the *biased* estimator given by MLE actually has a *lower* Mean Squared Error than the natural unbiased estimator. This illustrates the possibility of decreasing the MSE by trading off bias and variance.