

1 Introduction to Statistical Learning Theory

1.1 Learning Objectives

1. Identify the input, action, and outcome spaces for a given machine learning problem.
2. Provide an example for which the action space and outcome spaces are the same and one for which they are different.
3. Explain the relationships between the decision function, the loss function, the input space, the action space, and the outcome space.
4. Define the risk of a decision function and a Bayes decision function.
5. Provide example decision problems for which the Bayes risk is 0 and the Bayes risk is nonzero.
6. Know the Bayes decision functions for square loss and multiclass 0/1 loss.
7. Define the empirical risk for a decision function and the empirical risk minimizer.
8. Explain what a hypothesis space is, and how it can be used with constrained empirical risk minimization to control overfitting.

1.2 Concept Check Questions

1. Suppose $\mathcal{A} = \mathcal{Y} = \mathbb{R}$ and \mathcal{X} is some other set. Furthermore, assume $P_{\mathcal{X} \times \mathcal{Y}}$ is a discrete joint distribution. Compute a Bayes decision function when the loss function $\ell : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}$ is given by

$$\ell(a, y) = \mathbf{1}(a \neq y),$$

the 0 – 1 loss.

Solution. The Bayes decision function f^* satisfies

$$f^* = \arg \min_f R(f) = \arg \min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg \min_f P(f(X) \neq Y),$$

where $(X, Y) \sim P_{\mathcal{X} \times \mathcal{Y}}$. Let

$$f_1(x) = \arg \max_y P(Y = y \mid X = x),$$

the maximum a posteriori estimate of Y . If there is a tie, we choose any of the maximizers. If f_2 is another decision function we have

$$\begin{aligned} P(f_1(X) \neq Y) &= \sum_x P(f_1(x) \neq Y \mid X = x)P(X = x) \\ &= \sum_x (1 - P(f_1(x) = Y \mid X = x))P(X = x) \\ &\leq \sum_x (1 - P(f_2(x) = Y \mid X = x))P(X = x) \quad (\text{Defn of } f_1) \\ &= \sum_x P(f_2(x) \neq Y \mid X = x)P(X = x) \\ &= P(f_2(X) \neq Y). \end{aligned}$$

Thus $f^* = f_1$.

2. (\star) Suppose $\mathcal{A} = \mathcal{Y} = \mathbb{R}$, \mathcal{X} is some other set, and $\ell : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}$ is given by $\ell(a, y) = (a - y)^2$, the square error loss. What is the Bayes risk and how does it compare with the variance of Y ?

Solution. From Homework 1 we know that the Bayes decision function is given by $f^*(x) = \mathbb{E}[Y|X = x]$. Thus the Bayes risk is given by

$$\mathbb{E}[(f^*(X) - Y)^2] = \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] = \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y|X] - Y)^2|X]] = \mathbb{E}[\text{Var}(Y|X)],$$

where we applied the law of iterated expectations. The law of total variance states that

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}[\mathbb{E}(Y|X)].$$

This proves the Bayes risk satisfies

$$\mathbb{E}[\text{Var}(Y|X)] = \text{Var}(Y) - \text{Var}[\mathbb{E}(Y|X)] \leq \text{Var}(Y).$$

Recall from Homework 1 that $\text{Var}(Y)$ is the Bayes risk when we estimate Y without any input X . This shows that using X in our estimation reduces the Bayes risk, and that the improvement is measured by $\text{Var}[\mathbb{E}(Y|X)]$. As a sanity check, note that if X, Y are independent then $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ so $\text{Var}[\mathbb{E}(Y|X)] = 0$. If $X = Y$ then $\mathbb{E}(Y|X) = Y$ and $\text{Var}[\mathbb{E}(Y|X)] = \text{Var}(Y)$.

The prominent role of variance in our analysis above is due to the fact that we are using the square loss.

3. Let $\mathcal{X} = \{1, \dots, 10\}$, let $\mathcal{Y} = \{1, \dots, 10\}$, and let $A = \mathcal{Y}$. Suppose the data generating distribution, P , has marginal $X \sim \text{Unif}\{1, \dots, 10\}$ and conditional distribution $Y|X = x \sim \text{Unif}\{1, \dots, x\}$. For each loss function below give a Bayes decision function.

- (a) $\ell(a, y) = (a - y)^2$,
- (b) $\ell(a, y) = |a - y|$,
- (c) $\ell(a, y) = \mathbf{1}(a \neq y)$.

Solution.

- (a) From Homework 1 we know that $f^*(x) = \mathbb{E}[Y|X = x] = (x + 1)/2$.
- (b) From Homework 1, we know that $f^*(x)$ is the conditional median of Y given $X = x$. If x is odd, then $f^*(x) = (x + 1)/2$. If x is even, then we can choose any value in the interval

$$\left[\left\lfloor \frac{x+1}{2} \right\rfloor, \left\lceil \frac{x+1}{2} \right\rceil \right].$$

- (c) From question 1 above, we know that $f^*(x) = \arg \max_y P(Y = y|X = x)$. Thus we can choose any integer between 1 and x , inclusive, for $f^*(x)$.

4. Show that the empirical risk is an unbiased and consistent estimator of the Bayes risk. You may assume the Bayes risk is finite.

Solution. We assume a given loss function ℓ and an i.i.d. sample $(x_1, y_1), \dots, (x_n, y_n)$. To show it is unbiased, note that

$$\begin{aligned} \mathbb{E}[\hat{R}_n(f)] &= \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\ell(f(x_i), y_i)] \quad (\text{Linearity of } \mathbb{E}) \\ &= \mathbb{E}[\ell(f(x_1), y_1)] \quad (\text{i.i.d.}) \\ &= R(f). \end{aligned}$$

For consistency, we must show that as $n \rightarrow \infty$ we have $\hat{R}_n(f) \rightarrow R(f)$ with probability 1. Letting $z_i = \ell(f(x_i), y_i)$, we see that the z_i are i.i.d. with finite mean. Thus consistency follows by applying the strong law of large numbers.

5. Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = \mathcal{A} = \mathbb{R}$. Suppose you receive the (x, y) data points $(0, 5)$, $(.2, 3)$, $(.37, 4.2)$, $(.9, 3)$, $(1, 5)$. Throughout assume we are using the 0 – 1 loss.
- (a) Suppose we restrict our decision functions to the hypothesis space \mathcal{F}_1 of constant functions. Give a decision function that minimizes the empirical risk over \mathcal{F}_1 and the corresponding empirical risk. Is the empirical risk minimizing function unique?
 - (b) Suppose we restrict our decision functions to the hypothesis space \mathcal{F}_2 of piecewise-constant functions with at most 1 discontinuity. Give a decision function that minimizes the empirical risk over \mathcal{F}_2 and the corresponding empirical risk. Is the empirical risk minimizing function unique?

Solution.

- (a) We can let $\hat{f}(x) = 5$ or $\hat{f}(x) = 3$ and obtain the minimal empirical risk of 3/5. Thus the empirical risk minimizer is not unique.
 - (b) One solution is to let $\hat{f}(x) = 5$ for $x \in [0, .1]$ and $\hat{f}(x) = 3$ for $x \in (.1, 1]$ giving an empirical risk of 2/5. There are uncountably many empirical risk minimizers, so again we do not have uniqueness.
6. (★) Let $\mathcal{X} = [-10, 10]$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$ and suppose the data generating distribution has marginal distribution $X \sim \text{Unif}[-10, 10]$ and conditional distribution $Y|X = x \sim \mathcal{N}(a + bx, 1)$ for some fixed $a, b \in \mathbb{R}$. Suppose you are also given the following data points: $(0, 1)$, $(0, 2)$, $(1, 3)$, $(2.5, 3.1)$, $(-4, -2.1)$.
- (a) Assuming the 0 – 1 loss, what is the Bayes risk?
 - (b) Assuming the square error loss $\ell(a, y) = (a - y)^2$, what is the Bayes risk?
 - (c) Using the full hypothesis space of all (measurable) functions, what is the minimum achievable empirical risk for the square error loss?
 - (d) Using the hypothesis space of all affine functions (i.e., of the form $f(x) = cx + d$ for some $c, d \in \mathbb{R}$), what is the minimum achievable empirical risk for the square error loss?
 - (e) Using the hypothesis space of all quadratic functions (i.e., of the form $f(x) = cx^2 + dx + e$ for some $c, d, e \in \mathbb{R}$), what is the minimum achievable empirical risk for the square error loss?

Solution.

- (a) For any decision function f the risk is given by

$$\mathbb{E}[\mathbf{1}(f(X) \neq Y)] = P(f(X) \neq Y) = 1 - P(f(X) = Y) = 1.$$

To see this note that

$$P(f(X) = Y) = \frac{1}{20\sqrt{2\pi}} \int_{-10}^{10} \int_{-\infty}^{\infty} \mathbf{1}(f(x) = y) e^{-(y-a-bx)^2/2} dy dx = \frac{1}{20\sqrt{2\pi}} \int_{-10}^{10} 0 dx = 0.$$

Thus every decision function is a Bayes decision function, and the Bayes risk is 1.

- (b) By problem 2 above we know the Bayes risk is given by

$$\mathbb{E}[\text{Var}(Y|X)] = \mathbb{E}[1] = 1,$$

since $\text{Var}(Y|X = x) = 1$.

(c) We choose \hat{f} such that

$$\hat{f}(0) = 1.5, \hat{f}(1) = 3, \hat{f}(2.5) = 3.1, \hat{f}(-4) = 2.1,$$

and $\hat{f}(x) = 0$ otherwise. Then we achieve the minimum empirical risk of $1/10$.

(d) Letting

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2.5 \\ 1 & -4 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 3.1 \\ -2.1 \end{pmatrix}$$

we obtain (using a computer)

$$\hat{w} = \begin{pmatrix} \hat{d} \\ \hat{c} \end{pmatrix} = (A^T A)^{-1} A^T y = \begin{pmatrix} 1.4856 \\ 0.8556 \end{pmatrix}.$$

This gives

$$\hat{R}_5(\hat{f}) = \frac{1}{5} \|A\hat{w} - y\|_2^2 = 0.2473.$$

[Aside: In general, to solve systems like the one above on a computer you shouldn't actually invert the matrix $A^T A$, but use something like `w=A\backslash y` in Matlab which performs a QR factorization of A .]

(e) Letting

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2.5 & 6.25 \\ 1 & -4 & 16 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 3.1 \\ -2.1 \end{pmatrix}$$

we obtain (using a computer)

$$\hat{w} = \begin{pmatrix} \hat{e} \\ \hat{d} \\ \hat{c} \end{pmatrix} = (A^T A)^{-1} A^T y = \begin{pmatrix} 1.7175 \\ 0.7545 \\ -0.0521 \end{pmatrix}.$$

This gives

$$\hat{R}_5(\hat{f}) = \frac{1}{5} \|A\hat{w} - y\|_2^2 = 0.1928.$$

2 Excess Risk Decomposition, Stochastic Gradient Descent and Regularization

2.1 Topic 1: Excess Risk Decomposition

2.1.1 Learning Objectives

1. Give precise definitions for excess risk, approximation error, estimation error, and optimization error.
2. Suppose we have nested hypothesis spaces, say $\mathcal{H}_1 \subset \mathcal{H}_2$. Explain how we would expect the approximation error and estimation error to change when we change from \mathcal{H}_1 to \mathcal{H}_2 , all else fixed.
3. Explain how we would expect the approximation error and estimation error to change when we increase the sample size, all else fixed.
4. Explain optimization error, and write down an excess risk decomposition that incorporates approximation error, estimation error, and optimization error. Why might we have negative optimization error but never negative estimation error?

2.1.2 Concept Check Questions

1. Let $\mathcal{X} = \mathcal{Y} = \{1, 2, \dots, 10\}$, $\mathcal{A} = \{1, \dots, 10, 11\}$ and suppose the data distribution has marginal distribution $X \sim \text{Unif}\{1, \dots, 10\}$. Furthermore, assume $Y = X$ (i.e., Y always has the exact same value as X). In the questions below we use square loss function $\ell(a, x) = (a - x)^2$.
 - (a) What is the Bayes risk?
 - (b) What is the approximation error when using the hypothesis space of constant functions?
 - (c) Suppose we use the hypothesis space \mathcal{F} of affine functions.
 - i. What is the approximation error?
 - ii. Consider the function $\hat{f}(x) = x + 1$. Compute $R(\hat{f}) - R(f_{\mathcal{F}})$.

Solution.

- (a) The best decision function is $f^*(x) = x$. The associated risk is 0.
- (b) The best constant function is $f(x) = \mathbb{E}[Y] = \mathbb{E}[X] = 5.5$. This has risk

$$\mathbb{E}[(Y - 5.5)^2] = \text{Var}(Y) = \frac{33}{4},$$

by using (or deriving) the formula for the variance of a discrete uniform distribution. Thus the approximation error is $33/4$.

- (c)
 - i. The Bayes decision function is affine, so the approximation error is 0.
 - ii. The risk is

$$R(\hat{f}) = \mathbb{E}[(Y - \hat{f}(X))^2] = \mathbb{E}[(X - (X + 1))^2] = 1.$$

Thus the answer is 1.

2. (★) Let $\mathcal{X} = [-10, 10]$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$ and suppose the data distribution has marginal distribution $X \sim \text{Unif}(-10, 10)$ and $Y|X = x \sim \mathcal{N}(a + bx, 1)$. Throughout we assume the square loss function $\ell(a, x) = (a - x)^2$.
 - (a) What is the Bayes risk?
 - (b) What is the approximation error when using the hypothesis space of constant functions (in terms of a and b)?
 - (c) Suppose we use the hypothesis space of affine functions.
 - i. What is the approximation error?
 - ii. Suppose you have a fixed data set and compute the empirical risk minimizer $\hat{f}_n(x) = c + dx$. What is the estimation error (in terms of a, b, c, d) ?

Solution. Throughout we use the fact that $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

- (a) The best decision function is $f(x) = \mathbb{E}[Y|X = x] = a + bx$. This has risk

$$\mathbb{E}[(Y - a - bX)^2] = \mathbb{E}[\mathbb{E}[(Y - a - bX)^2|X]] = \mathbb{E}[1] = 1.$$

- (b) The best constant function is given by $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = a + b\mathbb{E}[X] = a$. This has risk

$$\mathbb{E}[(Y - a)^2] = \mathbb{E}[\mathbb{E}[(Y - a)^2|X]] = \mathbb{E}[1 + b^2X^2] = 1 + b^2\mathbb{E}[X^2],$$

where

$$\mathbb{E}[X^2] = \int_{-10}^{10} \frac{x^2}{20} dx = \frac{2000}{3 \cdot 20} = \frac{100}{3}.$$

Thus the approximation error is $100b^2/3$.

- (c) i. There is an affine Bayes decision function, so the approximation error is 0.
- ii. Note that

$$\begin{aligned} R(\hat{f}_n) &= \mathbb{E}[(Y - c - dX)^2] = \mathbb{E}[\mathbb{E}[(Y - c - dX)^2 | X]] \\ &= \mathbb{E}[1 + ((a - c) + (b - d)X)^2] = 1 + (a - c)^2 + 100(b - d)^2/3. \end{aligned}$$

Thus the estimation error is $(a - c)^2 + 100(b - d)^2/3$.

3. Try to best characterize each of the following in terms of one or more of optimization error, approximation error, and estimation error.
 - (a) Overfitting.
 - (b) Underfitting.
 - (c) Precise empirical risk minimization for your hypothesis space is computationally intractable.
 - (d) Not enough data.

Solution.

- (a) High estimation error due to insufficient data relative to the complexity of your hypothesis space. Can be accompanied by low approximation error indicating a complex hypothesis space.
 - (b) High approximation error due to an overly simplistic hypothesis space. Can be accompanied by low estimation error due to the large amount of data relative to the (low) complexity of the hypothesis space.
 - (c) Increased optimization error.
 - (d) High estimation error.
4. (a) We sometimes look at $R(\hat{f}_n)$ as random, and other times as deterministic. What causes this difference?
 - (b) True or False: Increasing the size of our hypothesis space can shift risk from approximation error to estimation error but always leaves the quantity $R(\hat{f}_n) - R(f^*)$ constant.
 - (c) True or False: Assume we treat our data set as a random sample and not a fixed quantity. Then the estimation error and the approximation error are random and not deterministic.
 - (d) True or False: The empirical risk of the ERM, $\hat{R}(\hat{f}_n)$, is an unbiased estimator of the risk of the ERM $R(\hat{f}_n)$.
 - (e) In each of the following situations, there is an implicit sample space in which the given expectation is computed. Give that space.
 - i. When we say the empirical risk $\hat{R}(f)$ is an unbiased estimator of the risk $R(f)$ (where f is independent of the training data used to compute the empirical risk).
 - ii. When we compute the expected empirical risk $\mathbb{E}[R(\hat{f}_n)]$ (i.e., the outer expectation).
 - iii. When we say the minibatch gradient is an unbiased estimator of the full training set gradient.

Solution.

- (a) The quantity is random when we consider the training data as a random sample of size n . If we focus on a fixed set of training data then the quantity is deterministic.
- (b) False. Note that \hat{f}_n depends on which hypothesis space you have chosen. As an example, imagine having an affine Bayes decision function, and changing the hypothesis space from the set of affine functions to the set of all decision functions. This can cause empirical risk minimization to overfit the training data thus creating a sharp rise in $R(\hat{f}_n) - R(f^*)$.
- (c) False, approximation error is a deterministic quantity.

- (d) False. The empirical risk of the ERM will often be biased low. This is why we use a test set to approximate its true risk. The issue is that \hat{f}_n depends on the training data so

$$\mathbb{E}\ell(\hat{f}_n(x_i), y_i) \neq \mathbb{E}\ell(\hat{f}_n(x), y)$$

where x, y is a new random draw from the data distribution that isn't in the training data.

- (e)
 - i. The space of training sets (i.e., samples of size n from the data generating distribution).
 - ii. The space of training sets (i.e., samples of size n from the data generating distribution).
 - iii. The space of all minibatches chosen from the full training set (i.e., samples of the batch size from the empirical distribution on the full training set).
5. For each, use \leq , \geq , or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subseteq \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ .
- (a) The estimation errors of two decision functions f_1, f_2 that minimize the empirical risk over the same hypothesis space, where f_2 uses 5 extra data points.
 - (b) The approximation errors of the two decision functions f_1, f_2 that minimize risk with respect to $\mathcal{F}_1, \mathcal{F}_2$, respectively (i.e., $f_1 = f_{\mathcal{F}_1}$ and $f_2 = f_{\mathcal{F}_2}$).
 - (c) The empirical risks of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively. Both use the same fixed training data.
 - (d) The estimation errors (for $\mathcal{F}_1, \mathcal{F}_2$, respectively) of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively.
 - (e) The risk of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively.

Solution.

- (a) Roughly speaking, more data is better, so we would tend to expect that f_2 will have lower estimation error. That said, this is not always the case, so the relationship cannot be determined.
 - (b) The approximation error of f_1 will be larger.
 - (c) The empirical risk of f_1 will be larger.
 - (d) Roughly speaking, increasing the hypothesis space should increase the estimation error since the approximation error will decrease, and we expect to need more data. That said, this is not always the case, so the answer is the relationship cannot be determined.
 - (e) Cannot be determined.
6. In the excess risk decomposition lecture, we introduced the decision tree classifier spaces \mathcal{F} (space of all decision trees) and \mathcal{F}_d (the space of decision trees of depth d) and went through some examples. The following questions are based on those slides. Recall that $P_{\mathcal{X}} = \text{Unif}([0, 1]^2)$, $\mathcal{Y} = \{\text{blue}, \text{orange}\}$, orange occurs with .9 probability below the line $y = x$ and blue occurs with .9 probability above the line $y = x$.
- (a) Prove that the Bayes error rate is 0.1.
 - (b) Is the Bayes decision function in \mathcal{F} ?
 - (c) For the hypothesis space \mathcal{F}_3 the slide states that $R(\tilde{f}) = 0.176 \pm .004$ for $n = 1024$. Assuming you had access to the training code that produces \tilde{f} from a set of data points, and random draws from the data generating distribution, give an algorithm (pseudocode) to compute (or estimate) the values 0.176 and .004.

Solution.

- (a) Since the output space is discrete and we are using the 0 – 1 loss, our best prediction is the highest probability output conditional on the input. By choosing orange below the line $y = x$ and blue above, we obtain a .1 probability of error. For the 0 – 1 loss, probability of error gives the risk.
- (b) No. Any decision tree in \mathcal{F} has finite depth, and thus will divide $[0, 1]^2$ into a finite number of rectangles. Thus we cannot produce the decision boundary $y = x$ used by the Bayes decision function.
- (c) Pseudocode follows:
 - i. Initialize L to be an empty list of risks.
 - ii. Repeat the following M times for some sufficiently large M :
 - A. Draw a random sample $(x_1, y_1), \dots, (x_n, y_n)$ from the data generating distribution.
 - B. Obtain a decision function \tilde{f} by running our training algorithm on the generated sample.
 - C. Draw a new random sample $(x'_1, y'_1), \dots, (x'_S, y'_S)$ of size S where S is sufficiently large.
 - D. Compute $e = |\{i \mid \tilde{f}(x'_i) \neq y'_i\}|$. That is, the number of times \tilde{f} is incorrect on our new sample.
 - E. Add e/S to the list L .
 - iii. Compute the sample average and standard deviation of the values in L . Above .176 would be the average and .004 would be the standard deviation.

Instead of drawing the sample of size S we could have computed the risk analytically.

2.2 Topic 2: Stochastic Gradient Descent

2.2.1 Learning Objectives

1. Be able to write the empirical risk for a particular loss function over a particular parameterized hypothesis space, such as for square loss over a hypothesis space of linear functions.
2. Compare and contrast gradient descent, minibatch gradient descent, and stochastic gradient descent.

2.2.2 Concept Check Questions

1. When performing mini-batch gradient descent, we often randomly choose the mini-batch from the full training set without replacement. Show that the resulting mini-batch gradient is an unbiased estimate of the gradient of the full training set. Here we assume each decision function f_w in our hypothesis space is determined by a parameter vector $w \in \mathbb{R}^d$.

Solution. Let $(x_{m_1}, y_{m_1}), \dots, (x_{m_N}, y_{m_N})$ be our mini-batch selected uniformly without replacement from the full training set $(x_1, y_1), \dots, (x_n, y_n)$.

$$\begin{aligned}
 \mathbb{E} \left[\nabla_w \frac{1}{N} \sum_{i=1}^N \ell(f_w(x_{m_i}, y_{m_i})) \right] &= \frac{1}{N} \sum_{i=1}^N \mathbb{E} [\nabla_w \ell(f_w(x_{m_i}, y_{m_i}))] && \text{(Linearity of } \nabla, \mathbb{E}) \\
 &= \frac{1}{N} \sum_{i=1}^N \mathbb{E} [\nabla_w \ell(f_w(x_{m_1}, y_{m_1}))] && \text{(Marginals are the same)} \\
 &= \mathbb{E} [\nabla_w \ell(f_w(x_{m_1}, y_{m_1}))] \\
 &= \sum_{i=1}^n \frac{1}{n} \nabla_w \ell(f_w(x_i, y_i)) \\
 &= \nabla_w \frac{1}{n} \sum_{i=1}^n \ell(f_w(x_i, y_i)) && \text{(Linearity of } \nabla).
 \end{aligned}$$

2. You want to estimate the average age of the people visiting your website. Over a fixed week we will receive a total of N visitors (which we will call our full population). Suppose the population mean μ is unknown but the variance σ^2 is known. Since we don't want to bother every visitor, we will ask a small sample what their ages are. How many visitors must we randomly sample so that our estimator $\hat{\mu}$ has variance at most $\epsilon > 0$?

Solution. Let x_1, \dots, x_n denote our randomly sampled ages, and let \hat{x} denote the sample mean $\frac{1}{n} \sum_{i=1}^n x_i$. Then

$$\text{Var}(\hat{x}) = \frac{\sigma^2}{n}.$$

Thus we require $n \geq \sigma^2/\epsilon$. Note that this doesn't depend on N , the full population size.

3. (★) Suppose you have been successfully running mini-batch gradient descent with a full training set size of 10^5 and a mini-batch size of 100. After receiving more data your full training set size increases to 10^9 . Give a heuristic argument as to why the mini-batch size need not increase even though we have 10000 times more data.

Solution. Throughout we assume our gradient lies in \mathbb{R}^d . Consider the empirical distribution on the full training set (i.e., each sample is chosen with probability $1/N$ where N is the full training set size). Assume this distribution has mean vector $\mu \in \mathbb{R}^d$ (the full-batch gradient) and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. By the central limit theorem the mini-batch gradient will be approximately normally distributed with mean μ and covariance $\frac{1}{n}\Sigma$, where n is the mini-batch size. As N grows the entries of Σ need not grow, and thus n need not grow. In fact, as N grows, the empirical mean and covariance matrix will converge to their true values. More precisely, the mean of the empirical distribution will converge to $\mathbb{E}\nabla\ell(f(X), Y)$ and the covariance will converge to

$$\mathbb{E}[(\nabla\ell(f(X), Y))(\nabla\ell(f(X), Y))^T] - \mathbb{E}[\nabla\ell(f(X), Y)]\mathbb{E}[\nabla\ell(f(X), Y)]^T$$

where $(X, Y) \sim P_{\mathcal{X} \times \mathcal{Y}}$.

The important takeaway here is that the size of the mini-batch is dependent on the speed of computation, and on the characteristics of the distribution of the gradients (such as the moments), and thus may vary independently of the size of the full training set.