Machine Learning and Computational Statistics
Hypothesis Space & Statistical Learning Theory

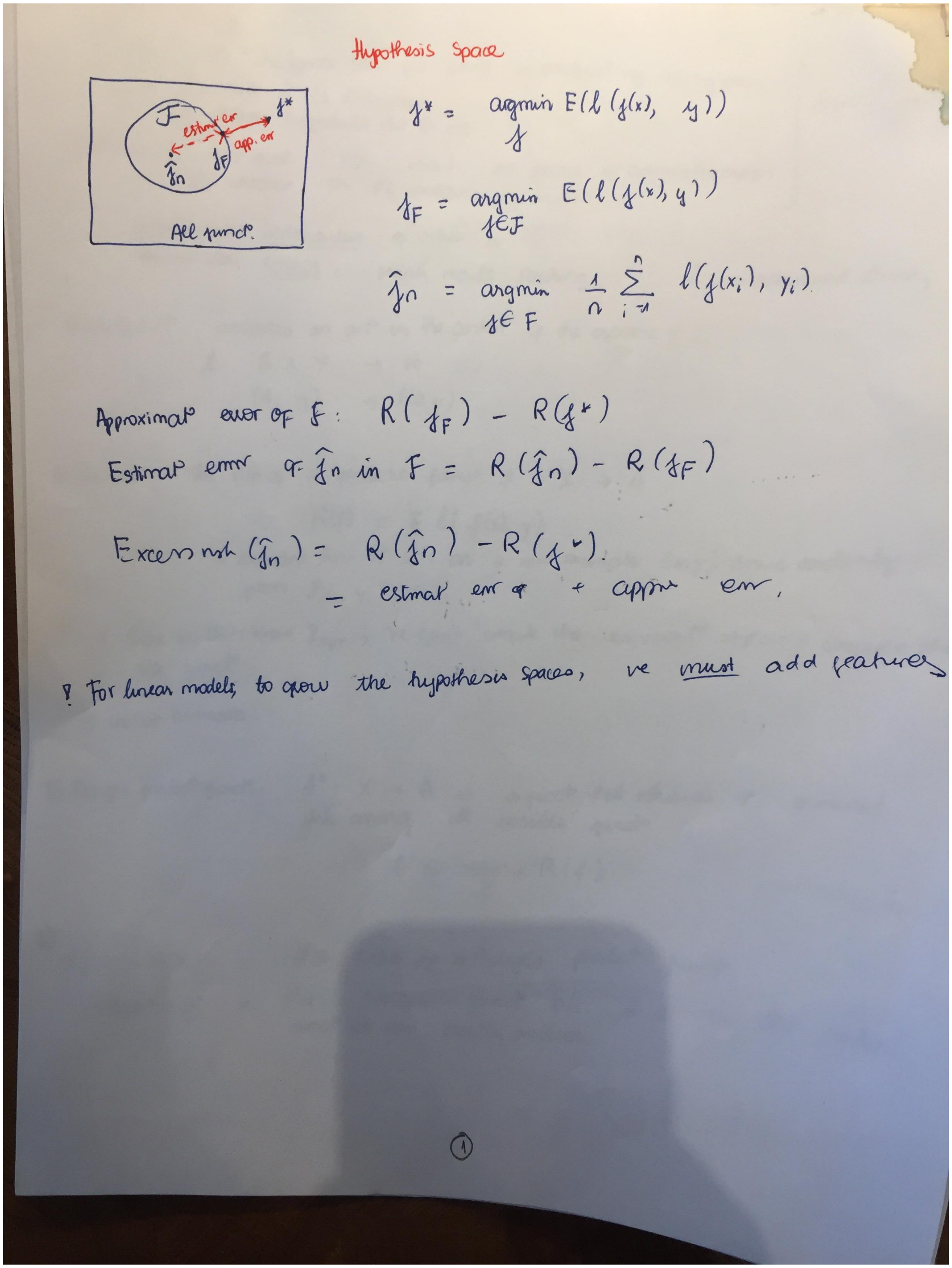
Note: This document consists of concepts and exercises related to hypothesis space and statistical learning theories.

1 Learning

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 Concepts



$\mathcal{F}$
estim° err
app. err
f^*
$\hat{f}_n$
f_F
All funct°
1

Hypothesis space

$$f^* = \underset{f}{\operatorname{argmin}}\, E(\ell(f(x), y))$$

$$f_F = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, E(\ell(f(x), y))$$

$$\hat{f}_n = \underset{f \in F}{\operatorname{argmin}}\, \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

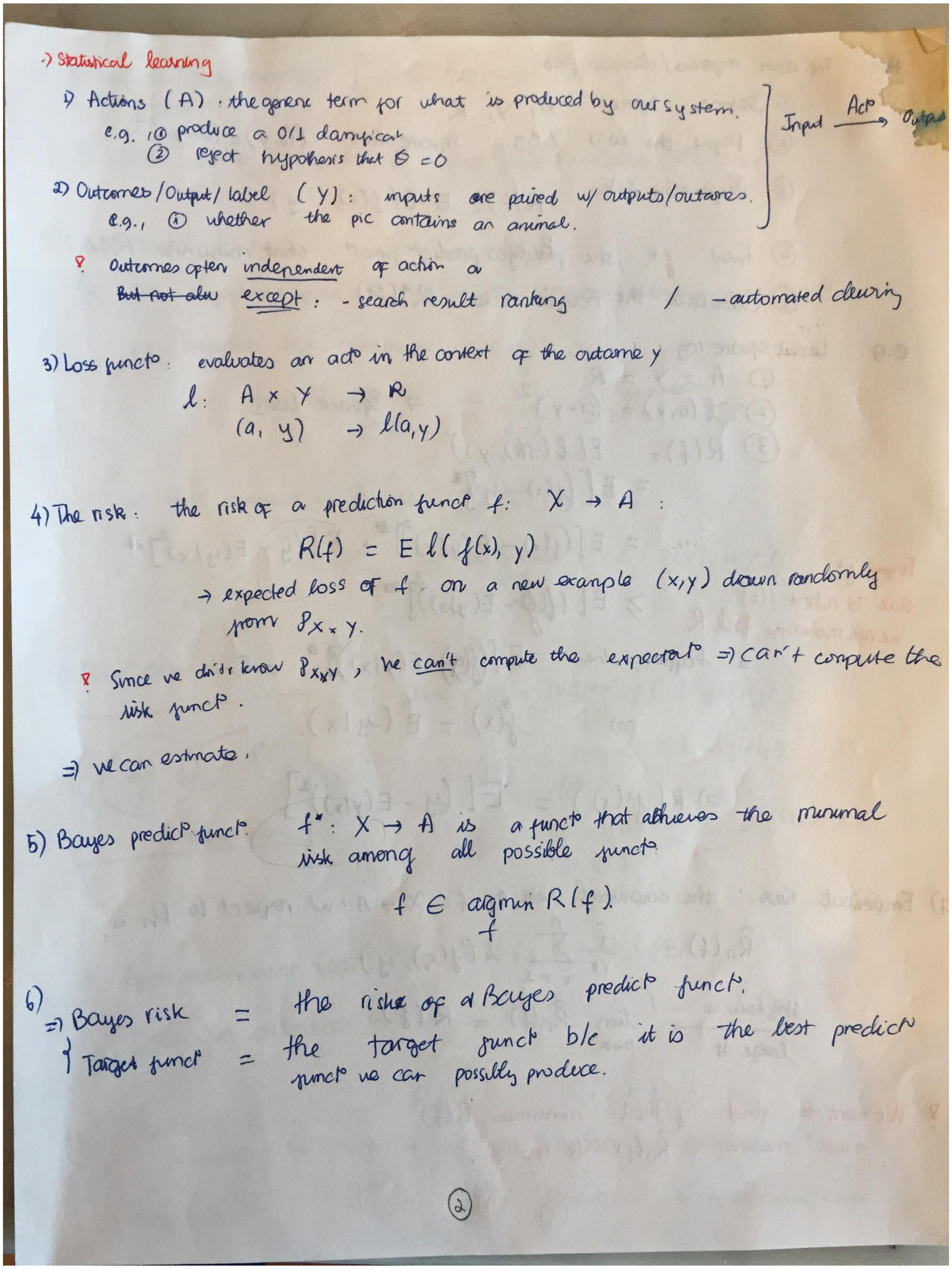
Approximat° error of $\mathcal{F}$: $R(f_F) - R(f^*)$

Estimat° error of $\hat{f}_n$ in $\mathcal{F}$ $= R(\hat{f}_n) - R(f_F)$

Excess risk $(\hat{f}_n) = R(\hat{f}_n) - R(f^*)$.

$=$ estimat° err of $+$ approx err,

! For linear models, to grow the hypothesis spaces, we <u>must</u> add features



-) Statistical learning

1) Actions (A) : the generic term for what is produced by our system.
e.g. 1① produce a 0/1 classificat°
② reject hypothesis that $\theta = 0$

2) Outcomes / Output / label (Y) : inputs are paired w/ outputs/outcomes.
e.g., ① whether the pic contains an animal.

Input $\xrightarrow{Act°}$ Output

? Outcomes often indenendent of action a
~~But not alw~~ except : - search result ranking / - automated driving

3) Loss funct° : evaluates an act° in the context of the outcome y

$$\ell : A \times Y \to \mathbb{R}$$
$$(a, y) \to \ell(a, y)$$

4) The risk : the risk of a prediction funct° $f : X \to A$:

$$R(f) = E\,\ell(f(x), y)$$

→ expected loss of f on a new example (x,y) drawn randomly from $P_{X \times Y}$.

? Since we didn't know $P_{X \times Y}$, we can't compute the expectat° ⇒ can't compute the risk funct°.

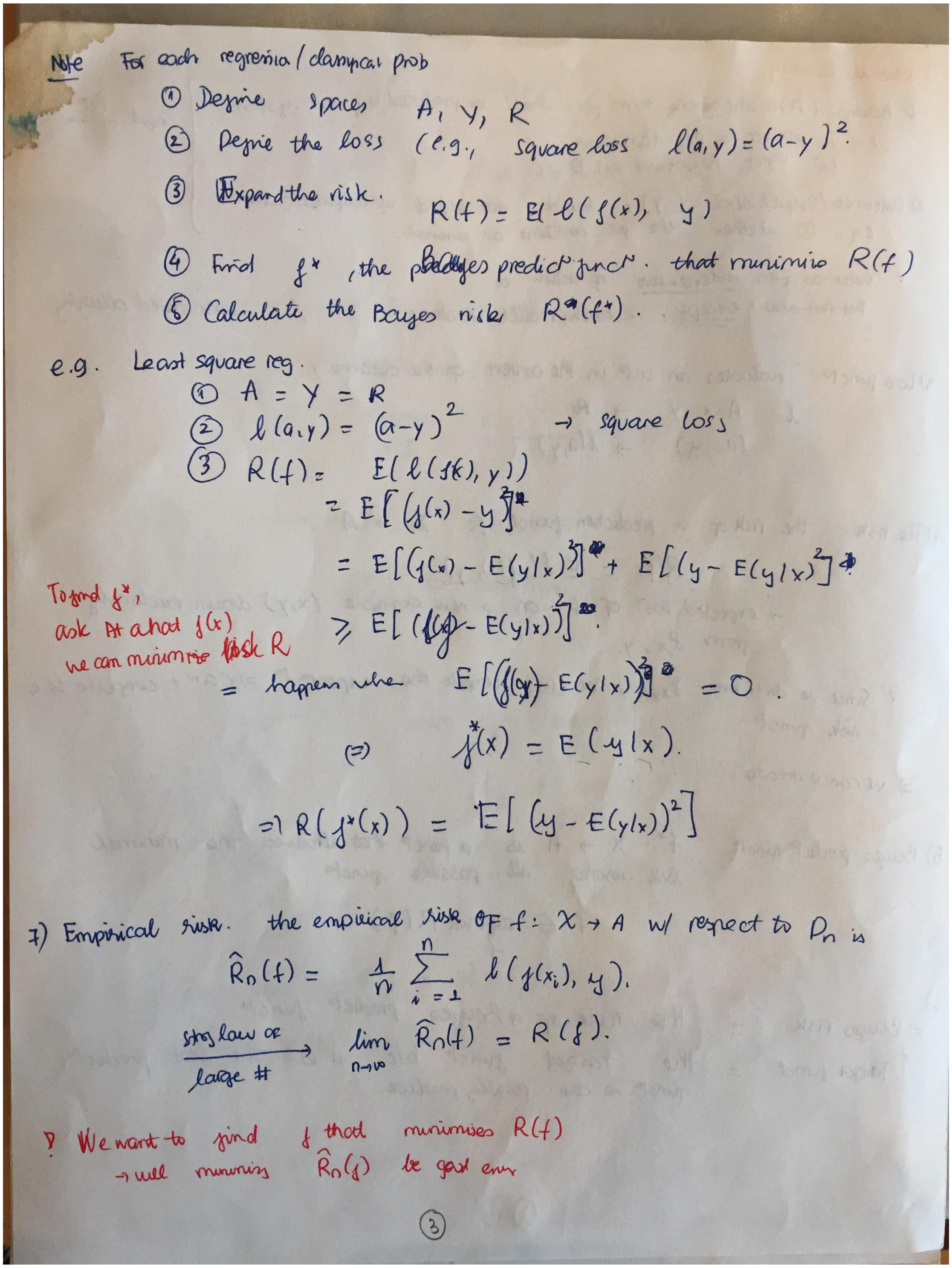
⇒ we can estimate.

5) Bayes predict° funct° $f^* : X \to A$ is a funct° that achieves the minimal risk among all possible funct°.

$$f \in \underset{f}{\operatorname{argmin}}\, R(f).$$

6)
⇒ Bayes risk = the risk of a Bayes predict° funct°.
} Target funct° = the target funct° b/c it is the best predict° funct° we can possibly produce.

②



Note For each regression / classifical prob

① Define spaces A, Y, R

② Define the loss (e.g., square loss $\ell(a,y) = (a-y)^2$.

③ ~~E~~Expand the risk. $R(f) = E\,\ell(f(x), y)$

④ Find f^*, the ~~pBdd~~ Bayes predicⁿ funcⁿ. that minimize $R(f)$

⑤ Calculate the Bayes risk $R^*(f^*)$.

e.g. Least square reg.

① $A = Y = R$

② $\ell(a,y) = (a-y)^2$ → square loss

③ $R(f) = E(\ell(f(x), y))$

$$= E[(f(x) - y)^2]$$

$$= E[(f(x) - E(y|x))^2] + E[(y - E(y|x))^2]$$

To find f^*,
ask at what $f(x)$
we can minimize risk R

$$\geq E[(f(x) - E(y|x))^2].$$

= happen when $E[(f(x) - E(y|x))^2] = 0$.

$$\Leftrightarrow \quad f^*(x) = E(y|x).$$

$$\Rightarrow R(f^*(x)) = E[(y - E(y|x))^2]$$

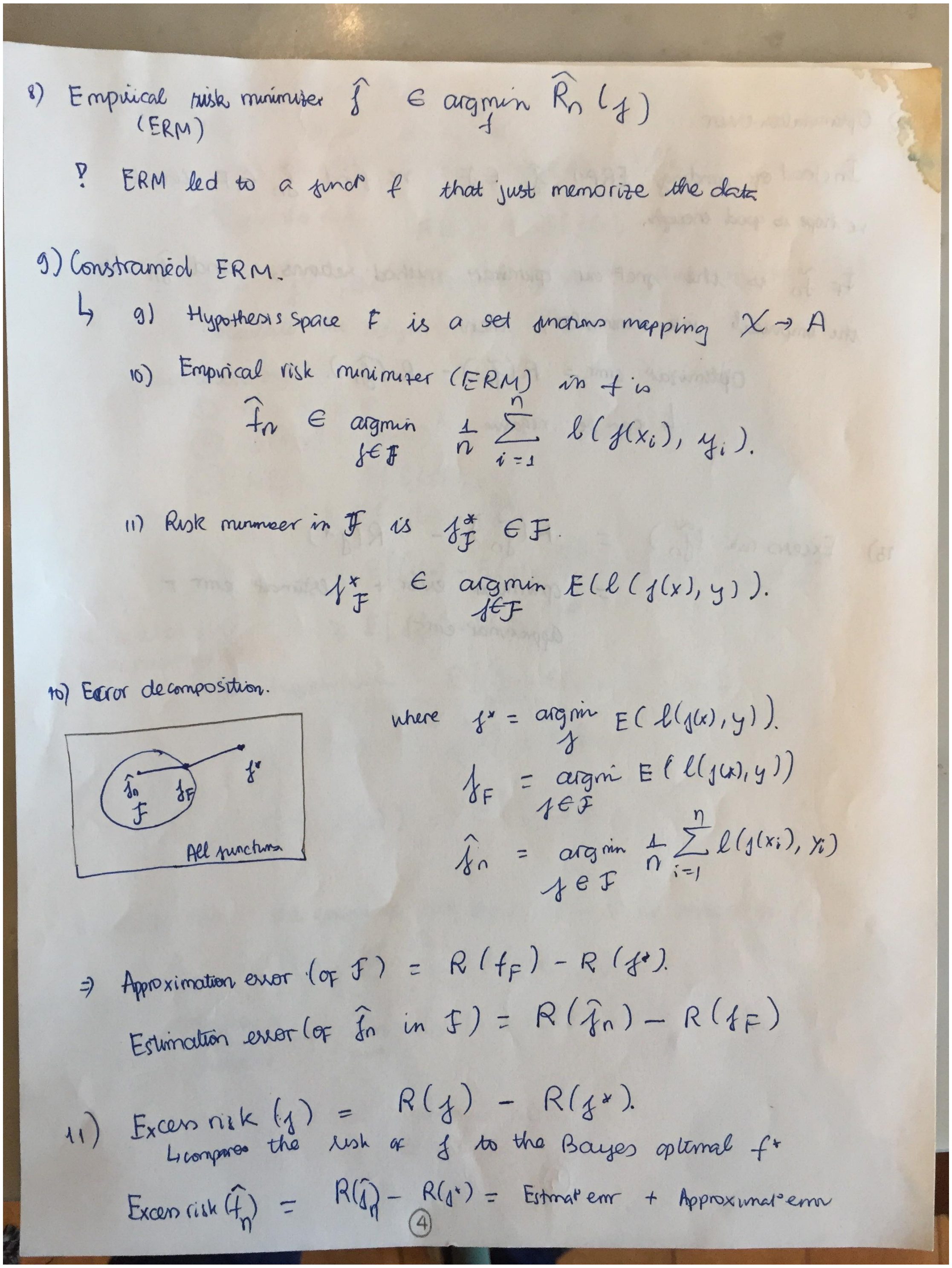
7) Empirical risk. the empirical risk of $f: \mathcal{X} \to A$ w/ respect to P_n is

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y).$$

$$\xrightarrow[\text{large \#}]{\text{strong law of}} \lim_{n \to \infty} \hat{R}_n(f) = R(f).$$

? We want to find f that minimises $R(f)$
→ will minimizing $\hat{R}_n(f)$ be good enough

③



8) Empirical risk minimizer $\hat{f} \in \underset{f}{\arg\min}\ \hat{R}_n(f)$
(ERM)

! ERM led to a fnct° f that just memorize the data

9) Constrained ERM.

↳ 9) Hypothesis space F is a set functions mapping $\mathcal{X} \to A$

10) Empirical risk minimizer (ERM) in f is

$$\hat{f}_n \in \underset{f\in\mathcal{F}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i).$$

11) Risk minimizer in $\mathcal{F}$ is $f^*_{\mathcal{F}} \in \mathcal{F}$.

$$f^*_{F} \in \underset{f\in F}{\arg\min}\ E(\ell(f(x), y)).$$

10) Error decomposition.

where $f^* = \underset{f}{\arg\min}\ E(\ell(f(x), y))$.

$$f_F = \underset{f\in F}{\arg\min}\ E(\ell(f(x), y))$$

$$\hat{f}_n = \underset{f\in F}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

⇒ Approximation error (of $\mathcal{F}$) $= R(f_F) - R(f^*)$.

Estimation error (of $\hat{f}_n$ in $\mathcal{F}$) $= R(\hat{f}_n) - R(f_F)$

11) Excess risk $(f) = R(f) - R(f^*)$.
↳ compares the risk of f to the Bayes optimal f^*

Excess risk $(\hat{f}_n) = R(\hat{f}_n) - R(f^*) =$ Estimat° err + Approximat° err

④

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.

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 $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ 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
$$\begin{array}{rcll} P(f_1(X) \neq Y) & = & \sum_x P(f_1(x) \neq Y|X = x)P(X = x) & \\ & = & \sum_x (1 - P(f_1(x) = Y|X = x))P(X = x) & \\ & \leq & \sum_x (1 - P(f_2(x) = Y|X = x))P(X = x) & \text{(Defn of } f_1\text{)} \\ & = & \sum_x P(f_2(x) \neq Y|X = x)P(X = x) & \\ & = & P(f_2(X) \neq Y). & \end{array}$$
Thus $f^* = f_1$.

2. ($\star$) Suppose $\mathcal{A} = \mathcal{Y} = \mathbb{R}$, $\mathcal{X}$ is some other set, and $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ 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
$$\mathbb{E}[(f^*(X) - Y)^2] = \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] = \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y|X] - Y)^2|X]] = \mathbb{E}[\mathrm{Var}(Y|X)],$$
where we applied the law of iterated expectations. The law of total variance states that
$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathbb{E}(Y|X)].$$

This proves the Bayes risk satisfies

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

Recall from Homework 1 that $\mathrm{Var}(Y)$ 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 $\mathrm{Var}[\mathbb{E}(Y|X)]$. As a sanity check, note that if X, Y are independent then $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ so $\mathrm{Var}[\mathbb{E}(Y|X)] = 0$. If $X = Y$ then $\mathbb{E}(Y|X) = Y$ and $\mathrm{Var}[\mathbb{E}(Y|X)] = \mathrm{Var}(Y)$.

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

3. Let $\mathcal{X} = \{1, \ldots, 10\}$, let $\mathcal{Y} = \{1, \ldots, 10\}$, and let $A = \mathcal{Y}$. Suppose the data generating distribution, P, has marginal $X \sim \mathrm{Unif}\{1, \ldots, 10\}$ and conditional distribution $Y|X = x \sim \mathrm{Unif}\{1, \ldots, x\}$. 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 $(x_1, y_i), \ldots, (x_n, y_n)$. To show it is unbiased, note that

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

For consistency, we must show that as $n \to \infty$ we have $\hat{R}_n(f) \to R(f)$ 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. ($\star$) 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

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

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

(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\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 Mathematical Fundamentals

The following questions are designed to check how prepared you are to take this class. Familiarity with linear algebra and probability at the level of these questions is expected for the class.

2.1 Probability

Let $(X_1, X_2, \cdots, X_d)$ have a d-dimensional multivariate Gaussian distribution, with mean vector $\mu \in \mathbf{R}^d$ and covariance matrix $\Sigma \in \mathbf{R}^{d \times d}$, i.e. $(X_1, X_2, \cdots, X_d) \sim \mathcal{N}(\mu, \Sigma)$ Use μ_i to denote the i^{th} element of μ and Σ_{ij} to denote the element at the i^{th} row and j^{th} column of Σ.

1. Let $x, y \in \mathbf{R}^d$ be two independent samples drawn from $\mathcal{N}(\mu, \Sigma)$. Give expression for $\mathbb{E}\|x\|_2^2$ and $\mathbb{E}\|x - y\|_2^2$. Express your answer as a function of μ and Σ. $\|x\|_2$ represents the $L2$-norm of vector x.

2. Find the distribution of $Z = \alpha_i X_i + \alpha_j X_j$, for $i \neq j$ and $1 \leq i, j \leq d$. The answer will belong to a familiar class of distribution. Report the answer by identifying this class of distribution and specifying the parameters.

3. (Optional) Assume W and R are two Gaussian distributed random variables. Is $W + R$ still Gaussian? Justify your answer.

2.2 Linear Algebra

1. Let A be a $d \times d$ matrix with rank k. Consider the set $S_A := \{x \in \mathbf{R}^d | Ax = 0\}$. What is the dimension of S_A?

2. Assume S_v is a k dimensional subspace in $\mathbf{R}^d$ and $v_1, v_2, \cdots, v_k$ form an orthonormal basis of S_v. Let w be an arbitrary vector in $\mathbf{R}^d$. Find

$$x^* = \underset{x \in S_v}{\text{argmin}} \|w - x\|_2,$$

where $\|w - x\|_2$ is the Euclidean distance between w and x. Express x^* as a function of $v_1, v_2, \ldots, v_k$ and w.

3. (Optional) Continuing from above, x^* can be expressed as

$$x^* = Mw,$$

where M is a $d \times d$ matrix. Prove that such an M always exists or more precisely find an expression for M as a function of $v_1, v_2, \cdots, v_k$. Compute the eigenvalues and one set of eigenvectors of M corresponding to the nonzero eigenvalues.

3 Risk Minimization

3.1 Square Loss

1. Let y be a random variable with a known distribution, and consider the square loss function $\ell(a, y) = (a-y)^2$. We want to find the action a^* that has minimal risk. That is, we want to find $a^* = \arg\min_a \mathbb{E}\,(a-y)^2$, where the expectation is with respect to y. Show that $a^* = \mathbb{E}y$, and the Bayes risk (i.e. the risk of a^*) is $\operatorname{Var}(y)$. In other words, if you want to try to predict the value of a random variable, the best you can do (for minimizing expected square loss) is to predict the mean of the distribution. Your expected loss for predicting the mean will be the variance of the distribution. [Hint: Recall that $\operatorname{Var}(y) = \mathbb{E}y^2 - (\mathbb{E}y)^2$.]

2. Now let's introduce an input. Recall that the **expected loss** or **"risk"** of a decision function $f : \mathcal{X} \to \mathcal{A}$ is
$$R(f) = \mathbb{E}\ell(f(x), y),$$
where $(x, y) \sim P_{\mathcal{X}\times\mathcal{Y}}$, and the **Bayes decision function** $f^* : \mathcal{X} \to \mathcal{A}$ is a function that achieves the *minimal risk* among all possible functions:
$$R(f^*) = \inf_f R(f).$$
Here we consider the regression setting, in which $\mathcal{A} = \mathcal{Y} = \mathbf{R}$. We will show for the square loss $\ell(a, y) = (a-y)^2$, the Bayes decision function is $f^*(x) = \mathbb{E}\left[y \mid x\right]$, where the expectation is over y. As before, we assume we know the data-generating distribution $P_{\mathcal{X}\times\mathcal{Y}}$.

    (a) We'll approach this problem by finding the optimal action for any given x. If somebody tells us x, we know that the corresponding y is coming from the conditional distribution $y \mid x$. For a particular x, what value should we predict (i.e. what action a should we produce) that has minimal expected loss? Express your answer as a decision function $f(x)$, which gives the best action for any given x. In mathematical notation, we're looking for $f^*(x) = \arg\min_a \mathbb{E}\left[(a-y)^2 \mid x\right]$, where the expectation is with respect to y. (Hint: There is really nothing to do here except write down the answer, based on the previous question. But make sure you understand what's happening...)

    (b) In the previous problem we produced a decision function $f^*(x)$ that minimized the risk for each x. In other words, for any other decision function $f(x)$, $f^*(x)$ is going to be at least as good as $f(x)$, for every single x. In math, we mean
$$\mathbb{E}\left[(f^*(x)-y)^2 \mid x\right] \le \mathbb{E}\left[(f(x)-y)^2 \mid x\right],$$
for all x. To show that $f^*(x)$ is the Bayes decision function, we need to show that
$$\mathbb{E}\left[(f^*(x)-y)^2\right] \le \mathbb{E}\left[(f(x)-y)^2\right]$$
for any f. Explain why this is true. (Hint: Law of iterated expectations.)

3.2 Median Loss

1. Show that for the absolute loss $\ell(\hat{y}, y) = |y - \hat{y}|$, $f^*(x)$ is a Bayes decision function if $f^*(x)$ is the median of the conditional distribution of y given x. [Hint: As in the previous section, consider one x at time. It may help to use the following characterization of a median: m is a median of the distribution for random variable y if $\mathbb{P}(y \geq m) \geq \frac{1}{2}$ and $\mathbb{P}(y \leq m) \geq \frac{1}{2}$.] Note: This loss function leads to "median regression". There are other loss functions that lead to "quantile regression" for any chosen quantile. (For partial credit, you may assume that the distribution of $y \mid x$ is discrete or continuous. For full credit, no assumptions about the distribution.)