



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2147*

Robust machine learning methods

MUHAMMAD OSAMA



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2022

ISSN 1651-6214
ISBN 978-91-513-1492-1
URN urn:nbn:se:uu:diva-472453

Dissertation presented at Uppsala University to be publicly examined in 101195, Ångström, Lägerhyddsvägen 1, Uppsala, Thursday, 9 June 2022 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Aalto Distinguished Professor Visa Koivunen (Aalto University, Finland).

Abstract

Osama, M. 2022. Robust machine learning methods. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2147. 50 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1492-1.

We are surrounded by data in our daily lives. The rent of our houses, the amount of electricity units consumed, the prices of different products at a supermarket, the daily temperature, our medicine prescriptions, our internet search history are all different forms of data. Data can be used in a wide range of applications. For example, one can use data to predict product prices in the future; to predict tomorrow's temperature; to recommend videos; or suggest better prescriptions. However in order to do the above, one is required to learn a model from data. A model is a mathematical description of how the phenomena we are interested in behaves e.g. how does the temperature vary? Is it periodic? What kinds of patterns does it have? Machine learning is about this process of learning models from data by building on disciplines such as statistics and optimization.

Learning models comes with many different challenges. Some challenges are related to how flexible the model is, some are related to the size of data, some are related to computational efficiency etc. One of the challenges is that of data outliers. For instance, due to war in a country exports could stop and there could be a sudden spike in prices of different products. This sudden jump in prices is an outlier or corruption to the normal situation and must be accounted for when learning the model. Another challenge could be that data is collected in one situation but the model is to be used in another situation. For example, one might have data on vaccine trials where the participants were mostly old people. But one might want to make a decision on whether to use the vaccine or not for the whole population that contains people of all age groups. So one must also account for this difference when learning models because the conclusion drawn may not be valid for the young people in the population. Yet another challenge could arise when data is collected from different sources or contexts. For example, a shopkeeper might have data on sales of paracetamol when there was flu and when there was no flu and she might want to decide how much paracetamol to stock for the next month. In this situation, it is difficult to know whether there will be a flu next month or not and so deciding on how much to stock is a challenge. This thesis tries to address these and other similar challenges.

In paper I, we address the challenge of data corruption i.e., learning models in a robust way when some fraction of the data is corrupted. In paper II, we apply the methodology of paper I to the problem of localization in wireless networks. Paper III addresses the challenge of estimating causal effect between an exposure and an outcome variable from spatially collected data (e.g. whether increasing number of police personnel in an area reduces number of crimes there). Paper IV addresses the challenge of learning improved decision policies e.g. which treatment to assign to which patient given past data on treatment assignments. In paper V, we look at the challenge of learning models when data is acquired from different contexts and the future context is unknown. In paper VI, we address the challenge of predicting count data across space e.g. number of crimes in an area and quantify its uncertainty. In paper VII, we address the challenge of learning models when data points arrive in a streaming fashion i.e., point by point. The proposed method enables online training and also yields some robustness properties.

Keywords: artificial intelligence, machine learning, risk minimization, data corruption, decision policy, conformal methods, data from contexts, online learning, sparse, robust, causal inference, point process, localization, distribution uncertainty, treatment rules, quantile treatment, predicting count data

Muhammad Osama, Department of Information Technology, Division of Systems and Control, Box 337, Uppsala University, SE-75105 Uppsala, Sweden. Department of Information Technology, Automatic control, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

© Muhammad Osama 2022

ISSN 1651-6214

ISBN 978-91-513-1492-1

URN urn:nbn:se:uu:diva-472453 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-472453>)

Dedicated to my parents and sister

Sammanfattning

Vi är omgivna av data i vårt dagliga liv. Hyran för våra hus, mängden elenheter som förbrukas, priserna på olika produkter i en stormarknad, dagstemperaturen, våra läkemedelsrecept, vår internet sökhistorik är alla olika former av data. Data är mycket användbart eftersom den kan användas för att göra många intressanta saker. Till exempel kan man använda data för att förutsäga produktpriser i framtiden, för att förutsäga morgondagens temperatur, för att rekommendera videor eller föreslå bättre läkemedelsrecept. Men för att göra ovanstående måste man lära sig en modell från data. En modell är en matematisk beskrivning av hur de fenomen vi är intresserade av beter sig. Till exempel hur beter sig temperaturen? Är det periodisk? Vilka typer av mönster har den? Maskininlärning handlar om att lära sig modeller från data genom att använda metoder från statistik och optimering.

Inlärning av modeller medför många utmaningar. Vissa utmaningar är relaterade till hur flexibel modellen är, vissa är relaterade till datastorlek, vissa är relaterade till beräkningseffektivitet etc. En av utmaningarna är extrema datapunkter. Till exempel, på grund av krig i ett land, kan data om priser på produkter i en stormarknad förvanstas. Den plötsliga prisstegringen är en avvikelse från den normala situationen och måste beaktas vid inlärning. En annan utmaning kan vara att data samlas in i en situation, men modellen ska användas i en annan situation. Till exempel kan man ha data från vaccinförsök där deltagarna mestadels var gamla människor, men man kanske vill dra en slutsats om vilket av vaccinerna som ska användas för hela befolkningen som innehåller människor i alla åldersgrupper. Så man måste också ta hänsyn till denna skillnad när man lär sig modeller eftersom slutsatsen annars kanske inte är giltig för de unga i befolkningen. Ytterligare en utmaning kan uppstå när data samlas in från olika källor eller sammanhang. Till exempel kan en butiksinnehavare ha uppgifter om försäljning av paracetamol när det var influensa och när det inte var influensa och hon kanske vill bestämma hur mycket paracetamol som ska lagras för nästa månad. I det här läget är det svårt att veta om det kommer att bli en influensa nästa månad eller inte, och det är därför en utmaning att bestämma sig för hur mycket hon ska lagra. Denna avhandling försöker ta itu med dessa och andra liknande utmaningar och föreslår lösningar för dem.

I artikel I tar vi upp utmaningen med datakorruption. Mer specifikt föreslår vi en metod som lär sig modeller på ett robust sätt även när en del av datan är korrupt. I artikel II tar vi upp utmaningen med datakorruption i trådlösa kommunikationsnätverk där målet är att lokalisera en trådlös nod. Data kan här bli korrupt ifall siktlinjen mellan noderna

är blockerad. Artikel III tar upp utmaningen att uppskatta orsakseffekten mellan en exponering- och en utfallsvariabel från rumsligt insamlad data (till exempel om ett ökand antal poliser i ett område minskar antalet brott där). Här kommer utmaningen på grund av dolda variabler som kan skapa samband mellan exponerings- och utfallsvariabler även om det inte finns något orsakssamband mellan dem. Paper IV tar upp utmaningen att lära sig förbättrade beslutspolicier, till exempel vilken behandling som ska tilldelas vilken patient givet tidigare data om behandlingar. I artikel V tittar vi på utmaningen med att lära sig modeller när data inhämtas från olika källor eller sammanhang och det framtida sammanhanget är okänt. I artikel VI tar vi upp utmaningen att förutsäga räkningsdata över ett område, t.ex. antal brott i en stad och kvantifiera osäkerheten i prediktionen med hjälp av giltiga prediktionsintervall. I artikel VII tar vi upp utmaningen med att lära modeller när datapunkter anländer punkt för punkt. Den föreslagna metoden möjliggör uppdatering av modellen online och ger även vissa robusthet egenskaper.

Acknowledgement

First of all, I want to thank Almighty Allah. He is the Creator, Designer and Sustainer of the universe. Everything is in His control. I thank Him for the life that He has given me and all the blessings in it. It would not have been possible for me to come to Sweden, do a PhD, make new friends and have this full-of-learning experience had He not willed for it. I pray that He makes me grateful and helps me live life in the way He has ordained.

I am also grateful to my supervisors Dave Zachariah and Thomas B. Schön. They have been extremely supportive throughout my PhD. Special thanks to Dave for he has been an absolutely supportive and wonderful PhD supervisor. He has guided me at every step. I pray that may he have all the success in life and may he and his family always stay blessed.

I would also like to thank Torbjörn Wigren and Jens Sjölund for their assistance in my job hunting process. Moreover, I would like to thank fellow colleagues at Systems and Control division for the friendly and congenial work environment. I wish everyone best of luck for their PhDs, careers and life in general.

Furthermore, I would like to thank Salman Toor from the division of scientific computing for his guidance in adjusting to the new environment of Sweden in my initial years of PhD.

In addition, I would like to thank my uncle Muhammad Arshad Jafer for his guidance during my school and college years. It was seeing him going abroad for higher studies that opened up this possibility in my mind.

Moreover, I would like to thank my parents for encouraging me for higher studies despite not being able to attend university themselves. I thank them for their support and the values that they have instilled in me. And yes I dare not forget my sister. Talking to her has always cheered me up whenever I have faced challenges or hardships while abroad.

Finally, I would also like to acknowledge the financial support of different funding agencies in my research: Swedish Research Council for project *NewLEADS - New Directions in Learning Dynamical Systems* (contract number: 621-2016-06079) and other projects (contract numbers: 2017-04610, 2018-05040, 2021-05022), Swedish Foundation for Strategic Research for project *ASSEMBLE* (contract number: RIT15-0012) and also *Wallenberg AI, Autonomous Systems and Software Program* (WASP).

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I M. Osama, D. Zachariah, and P. Stoica. “Robust risk minimization for statistical learning from corrupted data”. In: *IEEE Open Journal of Signal Processing* (2020), pp. 287–294
- II M. Osama, D. Zachariah, S. Dwivedi, and P. Stoica. “Robust localization in wireless networks from corrupted signals”. In: *EURASIP Journal on Advances in Signal Processing* 2021.1 (2021), pp. 1–18
- III M. Osama, D. Zachariah, and T. B. Schön. “Inferring heterogeneous causal effects in presence of spatial confounding”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4942–4950
- IV M. Osama, D. Zachariah, and P. Stoica. “Learning robust decision policies from observational data”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 18205–18214
- V M. Osama, D. Zachariah, and P. Stoica. “Robust learning in heterogeneous contexts”. In: *Operations Research Letters* (Submitted) (2022)
- VI M. Osama, D. Zachariah, and P. Stoica. “Prediction of spatial point processes: regularized method with out-of-sample guarantees”. In: *Advances in Neural Information Processing Systems* 32 (2019)
- VII M. Osama, D. Zachariah, P. Stoica, and T. B. Schön. “Online learning for prediction via covariance fitting: computation, performance and robustness”. In: *Journal of Machine Learning Research* (Submitted) (2022)

Reprints were made with permission from the publishers.

Contents

1	Overview	13
2	Introduction	17
2.1	Data generating process	17
2.2	Parameter set	21
2.3	Loss function	22
2.4	Best parameter	24
2.5	Empirical Risk Minimization	24
3	Limitations of Empirical Risk Minimization	29
3.1	Robustness against corrupted data	29
3.2	Robustness against confounding	33
3.3	Robustness against tail events	37
3.4	Robustness against uncertainty in distribution	40
3.5	Robustness against misspecification	44
	References	47

Paper I – Robust risk minimization for statistical learning from corrupted data

Paper II – Robust localization in wireless networks from corrupted signals

Paper III – Inferring heterogeneous causal effects in presence of spatial confounding

Paper IV – Learning robust decision policies from observational data

Paper V – Robust learning in heterogeneous contexts

Paper VI – Prediction of spatial point processes: regularized method with out-of-sample guarantees

Paper VII – Online learning for prediction via covariance fitting: computation, performance and robustness

1. Overview

Chapter 2 of the thesis gives an introduction to the basics of machine learning (ML) and builds towards a popular learning criterion used in ML; namely, empirical risk minimization. Subsequently, Chapter 3 highlights some limitations of empirical risk minimization and motivates different robust approaches developed in this thesis.

Below, a short summary and a statement of contribution for each individual paper included in the thesis is provided.

Statement of contribution

Paper I

This paper considers the problem of statistical learning when some fraction of training data is corrupted. Intuitively, the idea is to assign unknown weights to each data point in the training data and then jointly learn the weights and the model parameters. This is done by defining a risk with respect to a non-parametric distribution. Then it is proposed to minimize the risk jointly with respect to the non-parametric distribution and model parameters subject to some constraints. This gives robustness against corrupted points. A coordinate descent algorithm to solve the optimization problem is also proposed. The method is shown to perform well across wide variety of machine learning tasks.

Statement of contribution

The initial idea was proposed by Osama and further developed together with Zachariah. The experiments were designed together by Osama and Zachariah. The implementation was done by Osama. Osama and Zachariah were equally responsible for writing with important contribution and feedback from Stoica.

Paper II

This paper applies the method developed in Paper I to the problem of localization in a wireless network from timing based measurements, where the measurements could be corrupted due to non-line-of-sight conditions. The method is tested for different measurement schemes such

as time-of-arrival, time-difference of arrival etc. The method is shown to localize accurately even under non-line-of-sight conditions and different measurement schemes.

Statement of contribution

The idea to address this application was proposed by Zachariah and formulated together with Osama, who also implemented and tested the method. Osama and Zachariah were equally responsible for writing, with important contributions and feedback from Dwivedi and Stoica.

Paper III

This paper proposes a method for estimating spatially varying causal effect of an exposure variable (e.g. amount of fertilizer) on an outcome variable (e.g. crop yield) from data acquired over space when there are unobserved confounding variables. Intuitively, the idea to handle confounding is to use residuals of exposure and outcome instead of these quantities directly. However, these estimated residuals may themselves have errors and hence we propose a method that is robust to these errors. The method is tested on synthetic and real data.

Statement of contribution

The central ideas were formulated by Zachariah and developed together with Osama, who also implemented and tested the method. Osama and Zachariah were equally responsible for writing, with important contributions and feedback from Schön.

Paper IV

This paper addresses the problem of learning a new decision policy (e.g. which treatment to assign to a patient) from observational data on past decisions made in certain contexts (e.g. patient covariates) with associated costs (e.g. blood pressure increase). Unlike methods which focus on learning a policy that minimizes the *average* cost, we note that in safety-critical applications, such as medical decision support, it is of interest to learn robust policies which minimize the *tails* of the cost. We propose a method to learn a policy that minimizes the quantile of the cost and use conformal methodology to find a proxy for the quantile. The method is tested on synthetic and real data and shown to reduce the tails of the cost.

Statement of contribution

The central ideas were formulated by Zachariah and developed together with Osama, who also implemented and tested the method. Osama and

Zachariah were equally responsible for writing, with important contributions and feedback from Stoica.

Paper V

This paper addresses the problem of learning a decision parameter when data is obtained from different contexts and the future context is unknown. Intuitively, the idea is that when future context is unknown one could just learn a decision parameter that minimizes the average risk where average is taken over the context distribution. However, the estimate of context distribution is poor when number of context is large or number of data points is small. So one needs to be robust against that. Moreover, when data is obtained from different contexts, we note that using excess risk is a more appropriate criterion than risk. We build a robust method based on these motivations and show that it gives better worst-case performance as compared to other methods.

Statement of contribution

The central ideas were formulated by Zachariah and developed together with Osama, who also implemented and tested the method. Osama and Zachariah were equally responsible for writing, with important contributions and feedback from Stoica.

Paper VI

This paper is about building a model for predicting count data across space (e.g. number of crimes in a district) and using that to estimate the underlying intensity function. We also quantify uncertainty in the predicted count using conformal methodology. We use Poisson model and propose a regularized criterion that yields an out-of-sample guarantee and smaller prediction intervals empirically.

Statement of contribution

The central ideas were formulated by Zachariah and developed together with Osama, who also implemented and tested the method. Osama and Zachariah were equally responsible for writing, with important contributions and feedback from Stoica.

Paper VII

This paper addresses the problem of learning models when data points arrive in a streaming fashion. A linear smoother predictor and a covari-

ance model for the outcome is proposed. Furthermore, a covariance fitting criterion is specified to learn the hyperparameters of the covariance model. This yields a linear smoother predictor which can be updated online. It is also shown that the resulting predictor has several robustness properties.

Statement of contribution

The central ideas were formulated by Zachariah and developed together with Osama, who also implemented and tested the method. Osama and Zachariah were equally responsible for writing, with important contributions and feedback from Stoica and Schön.

2. Introduction

In the world today, many technologies and businesses are data-driven. For example, process industries use data to forecast maintenance of equipments to avoid losses arising from equipment breakdown. Producers of consumer goods and services use our search history on the Internet to show us targeted advertisements or recommend videos and movies. Banks and investment firms use data to assess risk of investments. All these examples involve statistical learning.

Statistical machine learning (ML) is a discipline which utilizes *data* to make *decisions*. Consider the following example. Suppose that a student applied to graduate school and receives offers from multiple research groups. The target here is to make a decision about which of the research groups to join. For this purpose, the student will look at data which will include factors such as research interest, supervisor, work of the supervisor's current PhD students, publications and future prospects, etc. He will weigh these factors in his mind and arrive at a decision that is best for him. In other words, he will be performing some sort of optimization in his head. Machine learning builds on different mathematical disciplines, such as statistics and optimization, to automatically learn decision rules from past data that will perform well in future.

2.1 Data generating process

The first important object in ML is the *data*, which we represent here by \mathcal{D} as a collection of n data points, i.e.,

$$\mathcal{D} = \{z_1, \dots, z_n\},$$

where z is a single data point. \mathcal{D} could be something as simple as recordings of monthly precipitation over the past n months or something as complicated as someone's internet search history for the past n days. We assume that the observed data arises from some data generating process (DGP). We describe this process by a probability distribution with a joint density function [3]

$$p(\mathcal{D}) = p(z_1, \dots, z_n). \tag{2.1}$$

Thus we assume that the observed data \mathcal{D} is just one possible draw from $p(\mathcal{D})$ and if we were to collect data again we would observe something different \mathcal{D}' .

In ML, it is often assumed that the data points are drawn in an independent and identically distributed (i.i.d.) manner. That is, the joint distribution $p(\mathcal{D})$ can be factored as

$$p(\mathcal{D}) = \prod_{i=1}^n p(\mathbf{z}_i). \quad (2.2)$$

Hence the n points that make up the data are n i.i.d. draws from $p(\mathbf{z})$. Let us look at a few examples of different $p(\mathbf{z})$.

Example 2.1 Suppose that an engineer wants to measure the temperature of a warehouse. She installs several temperature sensors to do that. Figure 2.1 shows the data obtained from the different temperature sensors. Here the data point z is a scalar. From the specification of the sensors, it is known that the errors of the sensors have a normal distribution. Hence the data distribution

$$p(z) = \mathcal{N}(z; \mu, \sigma^2)$$

is a Gaussian distribution where μ is equal to the underlying warehouse temperature. The goal in this example could be to estimate the underlying temperature from the obtained measurements.

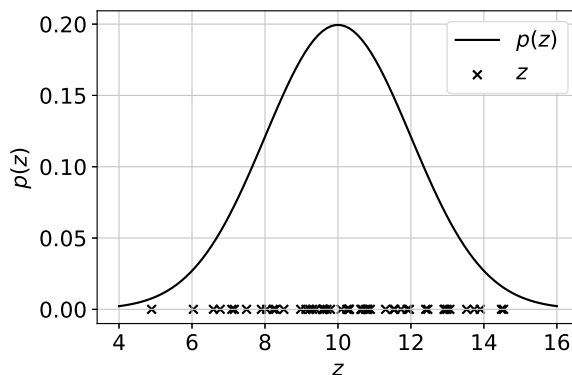


Figure 2.1. Data z on temperature (denoted by crosses) and density $p(z)$.

Example 2.2 Figure 2.2 shows data from students partaking in an ML course offered at a university. Here each data point is a pair, i.e., $\mathbf{z} = (x, y)$ where y represents the obtained marks of a student in the final exam of the course and x represents the hours of study spent preparing for the exam. Suppose the data is drawn i.i.d. from a Gaussian distribution [3], i.e.,

$$p(\mathbf{z}) = p(x, y) = \mathcal{N}(x, y; \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

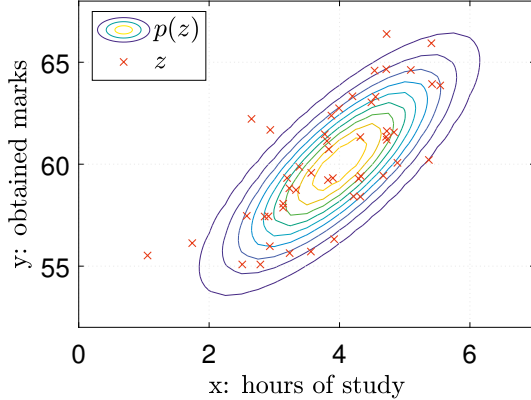


Figure 2.2. Data distribution $p(\mathbf{z})$, illustrated by contours, and the data points $\mathbf{z} = (x, y)$, denoted by red crosses, for regression data of Example 2.2

where $\boldsymbol{\mu} = [\mathbb{E}[x], \mathbb{E}[y]]^\top$ is the mean vector and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$ is the covariance matrix.

This is a typical example of many problems in ML where it is of interest to summarize the association between outcome y and vector of covariates \mathbf{x} .

Example 2.3 Figure 2.3 shows data on vaccination rate and mortality rate acquired over 2-dimensional space. Here each data point $\mathbf{z} = (\mathbf{s}, x, y)$ where $\mathbf{s} = (s_1, s_2)$ denotes the spatial coordinates, x is an exposure variable denoting the vaccination rate and y is an outcome variable denoting the mortality rate. The data is drawn i.i.d. from the following structural causal model (SCM) [31]:

$$\begin{aligned} \mathbf{s} &\sim \text{Unif}([0, 10]^2), \\ x &= f_x(\mathbf{s}, u_x) \in [0, 1], \\ y &= f_y(x, u_y) \in [0, 1]. \end{aligned} \tag{2.3}$$

An SCM is a set of equations that show how different variables are assigned values. For example, here the SCM describes that \mathbf{s} takes values uniformly at random in the region $[0, 10] \times [0, 10]$ in two dimensions, x takes values according to a function f_x of \mathbf{s} and some random noise u_x . And y is related to x according to another function f_y . The SCM in turn induces a data generating distribution $p(\mathbf{z}) = p(\mathbf{s}, x, y)$. A typical goal in these types of problems is to estimate average treatment effect [15] of the exposure x on the outcome y .

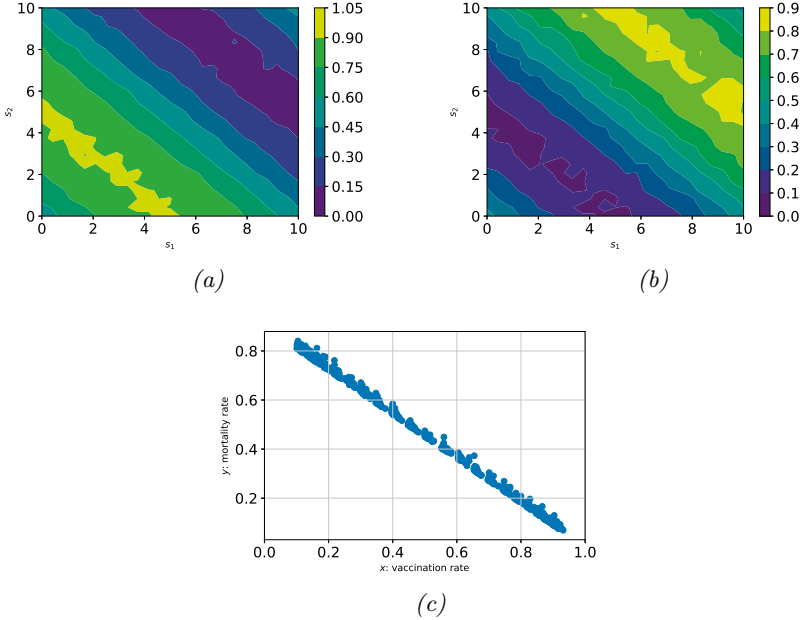


Figure 2.3. Data acquired over space: (a) Contour plot of vaccination rate x across space (b) Contour plot of mortality rate y across space (c) Scatter plot of x and y . Here the data generator is induced by the SCM in (2.3).

Example 2.4 Figure 2.4 shows data on blood pressure of female and male patients obtained as a result of using some policy to assign a treatment. Here each data point $\mathbf{z} = (t, x, y)$ where $t \in \{0, 1\}$ is a treatment variable indicating whether treatment was assigned or not. $x \in \{0, 1\}$ is a covariate denoting the gender of a patient and y is an outcome variable denoting the blood pressure. The data is drawn i.i.d. from the following data distribution that admits a causal factorization

$$p(t, x, y) = p(y|t, x)p(t|x)p(x). \quad (2.4)$$

Here $p(y|t, x) = \mathcal{N}(y; \mu(t, x), \sigma^2(t, x))$ is a normal distribution with mean and variance depending on t and x . $p(x = 1) = p(x = 0) = 0.5$ is the covariate distribution i.e. the probability of drawing a male ($x = 0$) or female ($x = 1$) patient from population. And $p(t|x)$ is the policy followed to assign treatment to a patient which is given as

$$p(t = 1|x) = \begin{cases} 0.6 & x = 1 \text{ (female)}, \\ 0.4 & x = 0 \text{ (male)}. \end{cases} \quad (2.5)$$

According to the above policy, female patients were assigned treatment with a higher probability than male patients. A typical goal in these kind

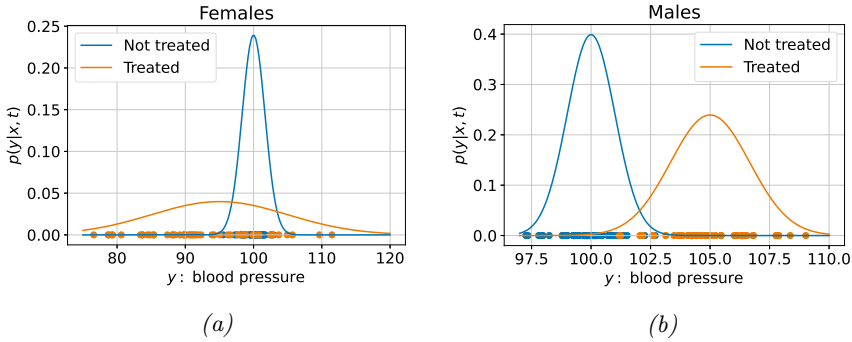


Figure 2.4. Plot of distribution $p(y|t, x)$ in (2.4) for gender covariate $x = \text{females}$ in (a) and for $x = \text{males}$ in (b). The orange and blue curves in (a) and (b) denote the distribution for the treated and untreated respectively. The orange and blue dots denote the data points for the treated and untreated respectively.

of problems is to improve upon the existing policy to get a better outcome for the population.

2.2 Parameter set

After data collection, the next important task is to make a *decision*. In Example 2.2, we might want to make a decision on what line describes the association between hours of study and obtained marks. A decision is indexed by a parameter θ and the set of possible decisions is denoted Θ , which we call the *parameter set*. Let us consider a few examples.

- In Example 2.1, if we are interested in fitting a Gaussian distribution to the data then

$$\Theta = \{\text{all Gaussians with a mean and variance}\} = \mathbb{R} \times \mathbb{R}^+.$$

- For Example 2.2, if we are interested in learning a straight line that summarizes the association between obtained marks and hours of study, we want to choose a line from a set of possible lines

$$\Theta = \{\text{all lines with a slope and intercept}\} = \mathbb{R}^2.$$

- For Example 2.3, if we are interested in learning the average causal effect of vaccination rate on death rate then

$$\Theta = \{\text{all possible values for ATE}\} = \mathbb{R}.$$

- For Example 2.4, if we are interested in whether a patient with covariate \mathbf{x} should be assigned treatment 1 or 0 then the set is all possible treatments i.e.,

$$\Theta = \{\text{all possible treatments}\} = \{0, 1\}.$$

In ML, we are interested in choosing a θ from Θ using data.

2.3 Loss function

When talking about choosing a target $\theta \in \Theta$, the first question that comes to mind is how should one make that choice? What should be the criterion? This brings us to the concept of a *loss function*.

The loss function quantifies the cost incurred in choosing a particular $\theta \in \Theta$ for a given data point $\mathbf{z} \sim p(\mathbf{z})$. Let us look at a few examples of some common loss functions used in ML.

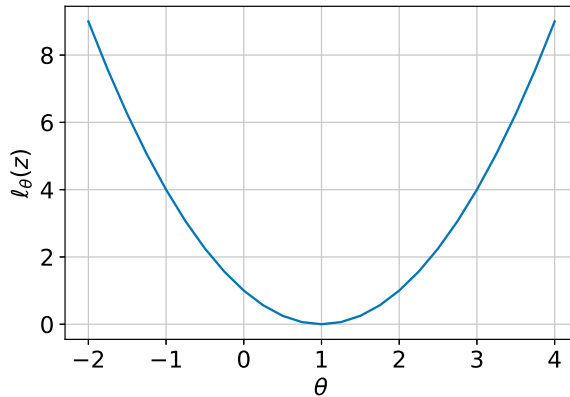


Figure 2.5. Illustration of squared loss in (2.6) evaluated at $z = 1$.

Example 2.5 One common loss function used in ML is the squared loss. For a scalar data point z as in Example 2.1, the squared loss is

$$\ell_{\theta}(z) = (z - \theta)^2. \quad (2.6)$$

Here $\theta \in \mathbb{R}$. Figure 2.5 shows the loss function against θ evaluated at a particular point.

Example 2.6 The squared loss for regression data in Example 2.2, where $\mathbf{z} = (x, y)$ and x is preparation time for exam and y is the obtained marks, is

$$\ell_{\theta}(x, y) = (y - x\theta_1 - \theta_0)^2. \quad (2.7)$$

Here $\theta \in \mathbb{R}^2$. Figure 2.6 shows the contour plot of the squared loss evaluated at a particular point.

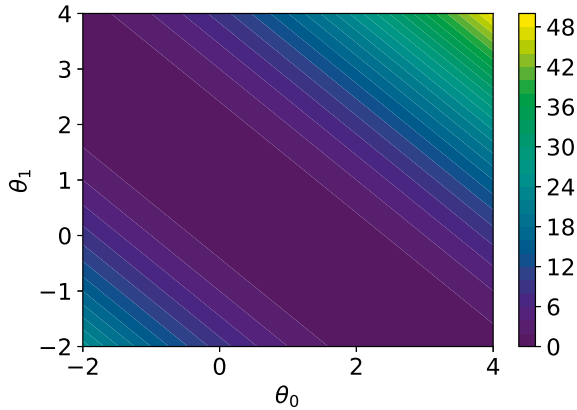


Figure 2.6. Contour plot of squared loss in (2.7) evaluated at point $(x, y) = (1, 1)$.

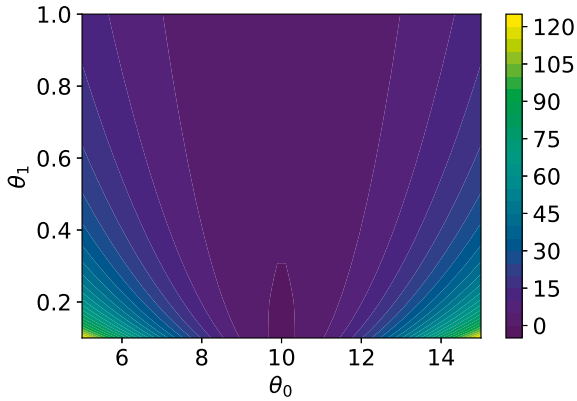


Figure 2.7. Contour plot of surprisal loss in (2.8) for point $z = 10$.

Example 2.7 Another commonly used loss function in ML is the surprisal or the negative log-likelihood loss [9]. For instance in Example 2.1, one may be interested in learning a parametric data model. Then the surprisal loss is

$$\ell_{\boldsymbol{\theta}}(\mathbf{z}) = -\ln p_{\boldsymbol{\theta}}(\mathbf{z}). \quad (2.8)$$

Intuitively, it quantifies the surprise of observing a data point \mathbf{z} under the model $p_{\boldsymbol{\theta}}(\mathbf{z})$. If \mathbf{z} is likely to come from $p_{\boldsymbol{\theta}}(\mathbf{z})$ then the loss is small. If the loss is very high then the data point \mathbf{z} is very surprising under the model $p_{\boldsymbol{\theta}}(\mathbf{z})$. Figure 2.7 shows the contour plot of surprisal loss evaluated at a particular point when $p_{\boldsymbol{\theta}}(\mathbf{z})$ is a one dimensional Gaussian distribution with mean θ_0 and variance θ_1 . Here $\boldsymbol{\theta} = [\theta_0, \theta_1]$.

2.4 Best parameter

Given the parameter set Θ and the loss function $\ell_{\theta}(\mathbf{z})$, we can define the optimal or best parameters as those which minimize the expected loss, i.e.,

$$\min_{\theta \in \Theta} \mathbb{E}_p[\ell_{\theta}(\mathbf{z})]. \quad (2.9)$$

In words, all the parameters that would give the minimum loss on average as we get different draws of data from $p(\mathbf{z})$. Under certain conditions, there is a unique parameter which minimizes (2.9) i.e.,

$$\theta^*(p) = \arg \min_{\theta \in \Theta} \mathbb{E}_p[\ell_{\theta}(\mathbf{z})], \quad (2.10)$$

for instance when the loss function is strictly convex in θ .

Example 2.8 Here we find the best parameter (2.10) for the data in Example 2.2 using the squared loss in Example 2.6. In this case, there exists a closed-form solution i.e.,

$$\theta^*(p) = \mathbb{E}_p \left[\begin{matrix} 1 & x \\ x & x^2 \end{matrix} \right]^{-1} \mathbb{E}_p \left[y \begin{bmatrix} 1 \\ x \end{bmatrix} \right]. \quad (2.11)$$

Figure 2.8 shows θ^* .

2.5 Empirical Risk Minimization

Note that $\theta^*(p)$ depends on the unknown distribution $p(\mathbf{z})$. We only observe n data points $\{\mathbf{z}_i\}_{i=1}^n$ from $p(\mathbf{z})$. In lieu of $p(\mathbf{z})$, we replace $p(\mathbf{z})$ by the empirical distribution constructed using the observed data points i.e.,

$$\hat{p}_n(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n \delta(\mathbf{z} - \mathbf{z}_i). \quad (2.12)$$

This gives us the following estimate

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell_{\theta}(\mathbf{z}_i), \quad (2.13)$$

which is called *empirical risk minimizer* (ERM) [40]. ERM is a very general principle for learning parameters in ML. For instance, ERM using surprisal loss $\ell_{\theta}(\mathbf{z}) = -\ln p_{\theta}(\mathbf{z})$ is equivalent to maximum likelihood principle [40]. Under certain boundedness conditions on the expected loss in (2.10), ERM is said to be *risk consistent* which means that the minimum value of the objective in (2.13) approaches the minimum value

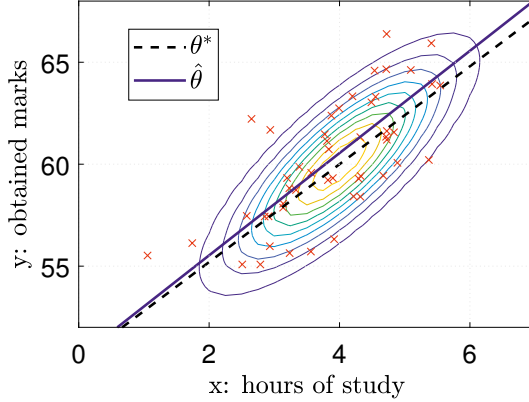


Figure 2.8. Best parameter θ^* and ERM estimate $\hat{\theta}$ for linear regression

of the objective in (2.10) as n increases [40]. Under regularity conditions [36, 44] $\hat{\theta}$ approaches θ^* as $n \rightarrow \infty$.

Closed-form solutions of ERM in (2.13) are not readily available and so one must resort to numerical search techniques. For low dimensional θ , one could use grid search techniques. However, for higher dimensions one must resort to gradient-based methods [22] which make use of the gradient of the objective in (2.13) i.e.,

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \eta \frac{1}{n} \sum_{i=1}^n \partial_{\theta} \ell_{\theta}(z_i) \Big|_{\hat{\theta}_t}. \quad (2.14)$$

In (2.14), one starts from an initial estimate $\hat{\theta}_0$ and then keeps updating by taking a step of size η in the direction opposite to the gradient until convergence. There are different gradient-based techniques for example Newton's based methods [22], majorization-minimization algorithms etc. Now let us see ERM in action in some examples.

Example 2.9 Here we find the ERM estimate (2.13) for the data in Example 2.2 using the squared loss in Example 2.6. ERM has a closed form solution in this case i.e.,

$$\begin{aligned} \hat{\theta} &= \mathbb{E}_n \begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix}^{-1} \mathbb{E}_n \begin{bmatrix} y \\ 1 \\ x \end{bmatrix}, \\ \hat{\theta} &= \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{y} \right). \end{aligned} \quad (2.15)$$

Here

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \mathbf{y} = [y_1, \dots, y_n]^{\top}.$$

Figure 2.8 shows $\hat{\theta}$ which is close to θ^* .

Example 2.10 Here we estimate the average treatment effect (ATE) τ of vaccination rate x on mortality rate y using data $\{x_i, y_i\}_{i=1}^n$ in Example 2.3. We use the ERM approach in (2.13) and squared loss with $\theta = [\theta_0, \theta_1]$ i.e.,

$$\hat{\theta} = \arg \min_{[\theta_0, \theta_1]} \frac{1}{n} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2. \quad (2.16)$$

Here θ_1 represents the ATE. The best and ERM estimates of ATE are

$$\tau^* = -0.9, \quad \hat{\tau} = -0.903. \quad (2.17)$$

Example 2.11 Suppose that we want to fit a Gaussian distribution $p_{\theta}(z) = \mathcal{N}(z; \mu, \sigma^2)$ to the temperature data in Example 2.1 where $\theta = [\mu, \sigma^2]$. We use the surprisal loss in (2.8). Then the best and ERM solutions are given as

$$(\mu^*, \sigma^*) = \arg \min_{\mu, \sigma} \mathbb{E}_p \left[\frac{1}{2} \log \sigma^2 + \frac{(z - \mu)^2}{2\sigma^2} \right], \quad (2.18)$$

and

$$(\hat{\mu}, \hat{\sigma}) = \arg \min_{\mu, \sigma} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} \log \sigma^2 + \frac{(z_i - \mu)^2}{2\sigma^2}. \quad (2.19)$$

The best and the ERM solutions are $[\mu^*, \sigma^*] = [10, 2]$ and $[\hat{\mu}, \hat{\sigma}] = [10.28, 2.25]$ respectively. Figure 2.9 shows the corresponding Gaussian distributions.

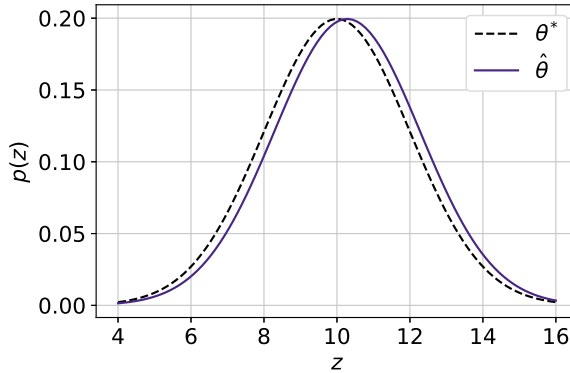


Figure 2.9. The best and ERM Gaussian distribution fit to temperature data in Example 2.1 using surprisal loss.

Example 2.12 Here we want to estimate a new policy for the data in Example 2.4 which if followed will reduce the average blood pressure. Let

$\theta \in \{0, 1\}$ be a binary decision policy denoting whether or not to assign the treatment. Then the best policy is given as

$$\theta^* = \arg \min_{\theta \in \{0,1\}} \mathbb{E}^\theta[y]. \quad (2.20)$$

Here the loss is simply the blood pressure itself. Moreover, \mathbb{E}^θ is the expectation with respect to distribution

$$p_\theta(t, x, y) = p(y|t, x)\mathbb{1}(t = \theta)p(x),$$

where $\mathbb{1}(t = \theta)$ is a degenerate distribution representing the new policy. We do not have data from p_θ and hence we can not evaluate the expectation in (2.20). However, \mathbb{E}^θ can be written in terms of expectation with respect to distribution $p(t, x, y)$ from which data is obtained in Example 2.4. Then the best policy becomes

$$\theta^* = \arg \min_{\theta \in \{0,1\}} \mathbb{E}[y|t = \theta]. \quad (2.21)$$

And the ERM estimate for the policy is given by

$$\hat{\theta} = \arg \min_{\theta \in \{0,1\}} \frac{1}{n_{t=\theta}} \sum_{i=1}^n y_i \mathbb{1}(t_i = \theta), \quad (2.22)$$

where $n_{t=\theta} = \sum_{i=1}^n \mathbb{1}(t_i = \theta)$. In words, the ERM policy is to choose the treatment which gives the lowest average blood pressure. Based on the distributions and data in Example 2.4, $\theta^* = \hat{\theta} = 1$ i.e., treat everyone.

3. Limitations of Empirical Risk Minimization

In the previous chapter, we discussed that empirical risk minimization is a general principle that, under some conditions, also has consistency properties as the number of data points increase. However, ERM has certain limitations and alternate *robust* methods are needed to address them. In this chapter, we look at some of these limitations and their corresponding solutions.

3.1 Robustness against corrupted data

In the best parameter $\theta^*(p)$ in (2.10), when $p(\mathbf{z})$ is replaced by the empirical distribution $\hat{p}_n(\mathbf{z})$ to obtain the ERM estimate in (2.13), it is inherently assumed that all observed samples $\{\mathbf{z}_i\}_{i=1}^n$ are drawn from $p(\mathbf{z})$. If this assumption is true then $\hat{p}_n(\mathbf{z})$ is a suitable approximation of $p(\mathbf{z})$. However, in practice some of the observed data \mathcal{D} could be corrupted by outliers, mislabeled examples or adversarial points. For example, when collecting data from several temperature sensors, data could be corrupted due to malfunctioning of some sensors. In such cases, it is more reasonable to assume that data is drawn from a *contaminated* distribution. One popular contamination model is the Huber contamination model [17] which assumes that data is drawn i.i.d. from the following distribution

$$\tilde{p}(\mathbf{z}) = (1 - \epsilon)p(\mathbf{z}) + \epsilon q(\mathbf{z}). \quad (3.1)$$

Here $\tilde{p}(\mathbf{z})$ is the observed distribution from which data points \mathbf{z}_i are drawn. It is a mixture of the target distribution $p(\mathbf{z})$ and a corrupting distribution $q(\mathbf{z})$. Furthermore, from the n data points that are observed, $1 - \epsilon$ fraction come from the target distribution and a small ϵ fraction come from the corrupting distribution $q(\mathbf{z})$. Under such contamination, the ERM in (2.13) would result in a poor estimate. Let us look at an example.

Example 3.1 We draw $n = 60$ data points $\mathbf{z} = (x, y)$ i.i.d from the following distribution

$$\tilde{p}(x, y) = (1 - \epsilon)p(x, y) + \epsilon q(x, y), \quad (3.2)$$

where

$$p(x, y) = \mathcal{N}\left(\begin{bmatrix} 4 \\ 60 \end{bmatrix}, \begin{bmatrix} 1 & 2.4 \\ 2.4 & 9 \end{bmatrix}\right), \quad (3.3)$$

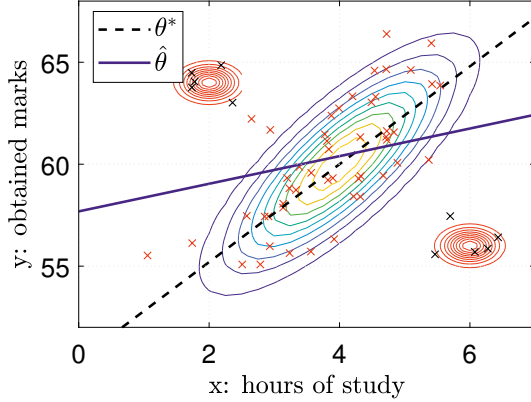


Figure 3.1. Linear regression example illustrating how ERM underperforms in the presence of corrupted data.

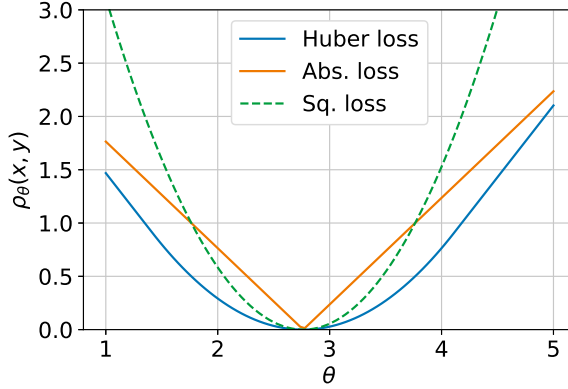


Figure 3.2. Plots of different loss functions. For the Huber loss in (3.6), $c = 1.345$ and $\sigma = 1$.

is the target distribution and

$$q(x, y) = 0.5\mathcal{N}\left(\begin{bmatrix} 6 \\ 56 \end{bmatrix}, \begin{bmatrix} 0.25^2 & 0 \\ 0 & 0.5^2 \end{bmatrix}\right) + 0.5\mathcal{N}\left(\begin{bmatrix} 2 \\ 64 \end{bmatrix}, \begin{bmatrix} 0.25^2 & 0 \\ 0 & 0.5^2 \end{bmatrix}\right) \quad (3.4)$$

is the corrupting distribution. The corruption fraction $\epsilon \approx 17\%$. Figure 3.1 shows the contours of $p(x, y)$ and $q(x, y)$ (in red). It also shows the best parameter θ^* (2.10) and the ERM estimate $\hat{\theta}$ (2.13). It can be seen that ERM performs poorly due to the corrupted data points.

Alternative solutions have been proposed in literature for estimation under the Huber contamination model. Classical approaches are build on using modified loss function $\rho_{\theta}(\mathbf{z})$. Here $\rho_{\theta}(\mathbf{z})$ is a *modified* or *less-*

sensitive loss function. In a sense, this approach is similar to (2.13) with $\ell_{\theta}(\mathbf{z})$ replaced by $\rho_{\theta}(\mathbf{z})$ and learning θ by solving

$$\arg \min_{\theta \in \Theta} \sum_{i=1}^n \rho_{\theta}(\mathbf{z}_i). \quad (3.5)$$

Consider the case of linear regression using the squared loss

$$\ell_{\theta}(\mathbf{z}) = (y - \mathbf{x}^{\top} \theta)^2,$$

where $\mathbf{z} = (\mathbf{x}, y)$. One example of $\rho_{\theta}(\mathbf{z})$ in this case is *absolute-deviation* loss

$$\rho_{\theta}(\mathbf{z}) = |y - \mathbf{x}^{\top} \theta|$$

This loss function is intuitively less sensitive to large errors as compared to the squared loss. Another example is *Huber's loss* [17, 47]

$$\rho_{\theta}(\mathbf{z}) = \frac{1}{2} \begin{cases} \frac{(y - \mathbf{x}^{\top} \theta)^2}{\sigma^2} & |y - \mathbf{x}^{\top} \theta| \leq c\sigma, \\ 2c \frac{|y - \mathbf{x}^{\top} \theta|}{\sigma} - c^2 & |y - \mathbf{x}^{\top} \theta| > c\sigma, \end{cases} \quad (3.6)$$

where σ is the standard deviation of the residual and c is a user-defined threshold. Huber's loss is also intuitively less sensitive to outliers: it behaves like a squared loss below the threshold c and like an absolute deviation loss above the threshold. Figure 3.2 shows the absolute deviation and the Huber loss compared with the squared loss.

Other methods in the literature make use of the same loss function i.e., $\ell_{\theta}(\mathbf{z})$ and aim to minimize the same objective as in (2.13). However, they either try to find robust estimates of the gradient of the objective or eliminate corrupted data points from the original data set. For example, the solution of (2.13) can be obtained via gradient updates of the following form

$$\theta^{t+1} = \theta^t - \eta \mathbb{E}_{\hat{\rho}_n}[\partial_{\theta} \ell_{\theta}(\mathbf{z})],$$

where the second term is the gradient of the objective in (2.13). When data is corrupted, the gradient above is poor. So some methods [32] try to come up with robust estimates of the gradient itself (using robust mean estimation methods [19]) which can then be plugged in the above gradient step.

In addition, some methods [11] try to remove outliers from the original data set \mathcal{D} to create a reduced data set $\tilde{\mathcal{D}}$. To identify and remove corrupted data points, they compute scores τ_i for each data point \mathbf{z}_i . Those data points whose scores are less than a certain threshold c are retained while others eliminated i.e.,

$$\tilde{\mathcal{D}} = \{\mathbf{z}_i\} \text{ s.t. } \tau_i \leq c.$$

Then ERM estimate is computed using \tilde{D} . This process is repeated until there are no more corrupted points identified. The scores and threshold make use of singular-value-decomposition of gradient matrix of the loss function, evaluated at the data points.

Another method is random sample consensus (RANSAC) [7] in which a random subset of the data containing a minimum number of data points is selected. Then a model is learned using the data points in the subset. All other data points in the training data are then tested against the learned model using a user-defined threshold to determine which data points are ‘inliers’ and become part of a *consensus* set. The model is improved by re-estimating it using points in the consensus set. This procedure is repeated a fixed number of times.

The above methods have several limitations. For methods that utilize some modified loss function, it might be easy to specify $\rho_{\theta}(\mathbf{z})$ for linear regression problems but it might not be straightforward to specify it for other machine learning problems such as classification etc. Moreover, methods which rely on robust gradient estimates or eliminating corrupted points from the data set, make use of scoring functions and user-specified thresholds (without clear guidelines) and also assume knowledge of the fraction of corrupted data ϵ . The sample consensus methods require sampling subsets containing at least as many data points as the number of parameters in the model which makes them unsuitable for high dimensional θ . Moreover, they also require specifying a user-defined threshold for identifying corrupted points.

In our work [28], we propose that since some fraction of the data is corrupted, it is not a good idea to replace $p(\mathbf{z})$ in (2.10) with $\hat{p}_n(\mathbf{z})$ in the first place. Instead, we propose to replace $p(\mathbf{z})$ by a non-parametric distribution with unknown probability weights i.e.,

$$p_{\pi}(\mathbf{z}) = \sum_{i=1}^n \pi_i \delta(\mathbf{z} - \mathbf{z}_i), \quad (3.7)$$

which gives us the following objective

$$\arg \min_{\theta, \pi} \mathbb{E}_{p_{\pi}} [\ell_{\theta}(\mathbf{z})]. \quad (3.8)$$

So intuitively the idea is to jointly minimize the *weighted loss* in (3.8) with respect to both θ and π . We additionally propose an entropy constraint on $p_{\pi}(\mathbf{z})$ to ensure that its support spans all the non-corrupted data points. Unlike above methods, our method does not require knowledge of the exact fraction of corruption and is free from scoring functions and user-defined thresholds. We also apply the method developed in [28] to the problem of localization in wireless networks when data is corrupted due to non-line-of-sight conditions [27].

3.2 Robustness against confounding

In Example 2.10, we were interested in the average treatment effect (ATE) of vaccination rate x on mortality rate y using ERM. In general, the ATE is defined as

$$\tau \triangleq \frac{\partial}{\partial \tilde{x}} \tilde{\mathbb{E}}[y|\tilde{x}], \quad (3.9)$$

where $\tilde{\mathbb{E}}[y|\tilde{x}]$ is the conditional outcome y were the exposure x intervened upon and set to \tilde{x} . The expectation can also be written as

$$\tilde{\mathbb{E}}[y|\tilde{x}] = \tilde{\mathbb{E}}[y], \quad (3.10)$$

where $\tilde{\mathbb{E}}[\cdot]$ is the expectation with respect to the following distribution

$$\tilde{p}(c, x, y) = p(y|x, c)\mathbb{1}(x = \tilde{x})p(c). \quad (3.11)$$

Here c represents confounders which are common causes of both exposure x and outcome y . For instance in Example 2.10, income could be a confounder since it might affect both who can afford the vaccine and also who has a better overall health. Our target is to find the ATE in (3.9), however, the problem is that in practice we do not observe data from (3.11) and hence cannot evaluate the expectation in (3.9). Instead, data is observed from

$$p(c, x, y) = p(y|x, c)p(x|c)p(c), \quad (3.12)$$

which admits a causal factorization according to the directed acyclic graph (DAG) in Figure 3.3a. Note that contrary to (3.11), where the exposure is fixed at \tilde{x} , the assignment of x depends on confounder c according to $p(x|c)$ in (3.12). The DAG corresponding to the distribution in (3.11) is shown in Figure 3.3b.

It is still possible to estimate the ATE by re-writing the expectation $\tilde{\mathbb{E}}[y]$ in the following way

$$\tilde{\mathbb{E}}[y] = \mathbb{E} \left[\frac{\tilde{p}(c, x, y)}{p(c, x, y)} y \right] = \mathbb{E} \left[\frac{\mathbb{1}(x = \tilde{x})}{p(x|c)} y \right] \neq \mathbb{E}[y|x = \tilde{x}], \quad (3.13)$$

where $\mathbb{E}[\cdot]$ is expectation with respect to distribution in (3.12) from which data is observed. Note that to be able to re-write the expectation as above we assume that $p(x|c) > 0$ for all x and c . (3.13) reveals some important points. First it shows that the expected outcome were the exposure fixed to \tilde{x} is not the same as the outcome condition on \tilde{x} . So for instance, if x were a binary treatment variable and we were interested in finding whether assigning treatment improves blood pressure y , we cannot simply compare the blood pressures of those treated to those untreated in the observed data. Instead, (3.13) shows that one needs to

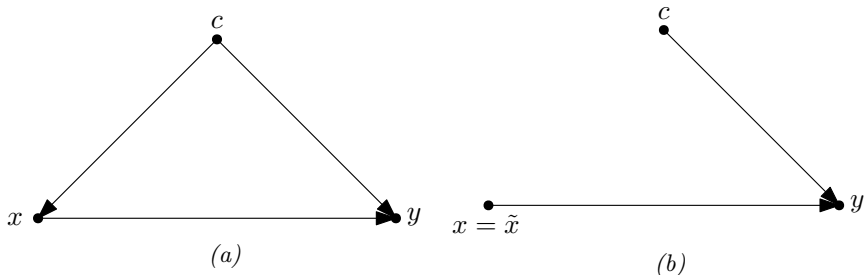


Figure 3.3. (a) Direct acyclic graph (DAG) showing relationship between exposure x , outcome y and confounder c . It induces distribution in (3.12) (b) DAG where the exposure is set to \tilde{x} . It induces distribution in (3.11).

re-weight the observed outcome by $p(x|c)$ to adjust for the confounding variable. This shows us how the presence of confounders must be taken into account to estimate the ATE.

This technique of re-weighting the outcome as above to adjust for confounders is called *inverse probability weighting* [13]. Another technique is called *regression adjustment* which also comes from (3.13). Replacing $p(x|c)$ by $\frac{p(c|x)p(x)}{p(c)}$ from Baye's rule in (3.13) we get

$$\tilde{\mathbb{E}}[y] = \mathbb{E}_c [\mathbb{E}[y|\tilde{x}, c]]. \quad (3.14)$$

Here \mathbb{E}_c represents expectation with respect to $p(c)$. So in regression adjustment one can estimate ATE by marginalizing $\mathbb{E}[y|\tilde{x}, c]$ over c .

When confounding variables are observed then one can adjust for them using the techniques indicated above. However, it may be that some or all confounding variables are hidden. In that case, a direct estimation using ERM and observed data on only x and y would be problematic. Let us look at an example.

Example 3.2 We modify the SCM of Example 2.3 as

$$\begin{aligned} \mathbf{s} &\sim \text{Unif}([0, 10]^2), & c &= f_c(\mathbf{s}, u_c), \\ x &= f_x(c, u_x), & y &= f_y(c, u_y). \end{aligned} \quad (3.15)$$

Here c represents standardized income which acts as a confounder and it affects both vaccination rate x and mortality rate y . We assume that c is unobserved. Figure 3.4 shows the plots of different variables across space. Note that here the exposure x has no direct effect on outcome y i.e., ATE $\tau = 0$. However, they are still correlated across space and using ERM gives us an ATE estimate of $\hat{\tau} = 0.812$. This illustrates the problem of confounding.

The problem of confounding when it occurs in the context of data collected over space is called *spatial confounding* and there are different

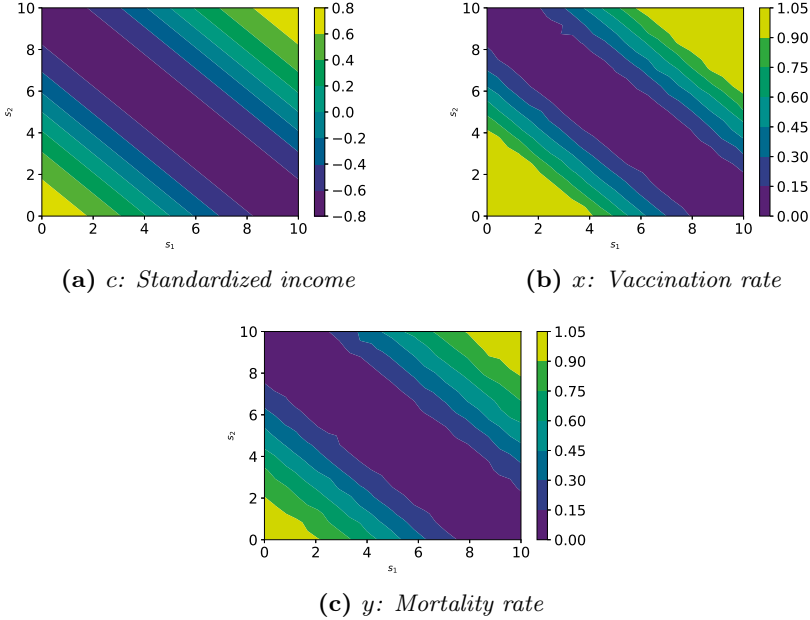


Figure 3.4. Illustration of spatial confounding: (a) Contour plot of confounder c denoting standardized income across space (b) Contour plot of exposure vaccination rate x across space (c) Contour plot of outcome mortality rate y across space. Note that y and x are correlated across space although ATE is zero here.

approaches in literature to handle spatial confounding even when c is unobserved.

Many approaches in literature use a linear model to estimate the ATE i.e.,

$$y = \theta_0 + \theta_1 x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Here the noise is assumed to be i.i.d normal. However, if there is a hidden confounding variable that varies spatially, the noise would have a spatial structure. For instance, if the data were generated linearly as

$$y = \mu_0 + \tau x + \underbrace{\beta c(s)}_{\epsilon(s)} + u_y \tag{3.16}$$

then the noise would be spatially varying (here $c(s)$ denotes a spatially varying confounder). One approach is to allow for spatial structure in the noise, an approach called *spatial generalized linear mixed model* [2]. i.e.,

$$y = \theta_0 + \theta_1 x + \tilde{\epsilon}(s)$$

where $\tilde{\epsilon}(s)$ is a spatially correlated noise. There are different ways of specifying a spatial structure for $\tilde{\epsilon}(s)$. For instance, one could specify

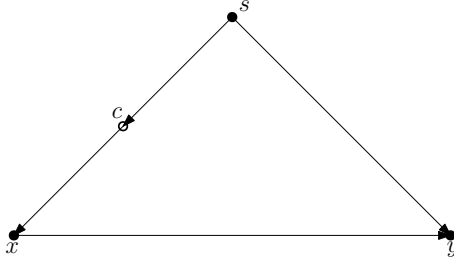


Figure 3.5. DAG with unobserved counfounder c . Spatial location s can be used to block the backdoor path between x and y .

that given n data points the $n \times 1$ vector of noise $\tilde{\epsilon}$ is jointly Gaussian with some covariance matrix i.e.,

$$\tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma).$$

Other approaches use some spatial basis function [10] i.e.,

$$\tilde{\epsilon}(s) = \phi^\top(s)\eta, \quad \eta \sim \mathcal{N}(\mathbf{0}, \mathbf{K}).$$

Although a spatially structured noise allows us to account for the unobserved spatially varying confounder, we still have the problem that this structured noise $\tilde{\epsilon}$ is correlated with the spatially varying exposure x . And the estimate of ATE may still have bias [16, 29]. To handle this problem, the idea is to allow for spatial structure in the noise while also ensuring that it is uncorrelated to x i.e.,

$$y = \theta_0 + \theta_1 x + \epsilon_\perp(s),$$

where

$$\epsilon_\perp(s) \perp x.$$

This approach is called *restricted spatial regression* [14, 16, 18]. To ensure the above, the vector of spatially structured noise $\tilde{\epsilon}$ for observed data is projected onto the space orthogonal to the span of the vector of exposure \mathbf{x} i.e.,

$$\epsilon_\perp = (\mathbf{I} - \mathbf{P}_x)\tilde{\epsilon}, \quad \tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

where \mathbf{P}_x is the projection matrix on the span of \mathbf{x} .

In our work [23], we take a different approach which blocks the *non-causal association* between x and y . We assume that data follows causal structure in DAG in Figure 3.5. Then we try to block the non-causal association between x and y by conditioning on space s . We note that even if the confounding variable c is unobserved, under the DAG in Figure 3.5, conditioning on s would block what is known as the *backdoor path* [31] between exposure x and outcome y . In which case, we have

$$\tilde{\mathbb{E}}[y|s] = \mathbb{E}[y|x = \tilde{x}, s].$$

So one can build a model for $\mathbb{E}[y|x = \tilde{x}, s]$ to estimate the causal effect. We assume a linear model as in (3.16). Moreover, we note that if one defines residuals $w = y - \mathbb{E}[y|s]$ and $v = x - \mathbb{E}[x|s]$ then

$$w = \tau v + \epsilon.$$

In other words, one can estimate the ATE from the residuals w and v . This avoids the need for specifying a spatial structure on noise and the problem of it being correlated with the exposure variable as discussed above.

3.3 Robustness against tail events

In Example 2.12, we made a single decision of treating all patients irrespective of whether the patient was a male or a female. However in medicine, people with different covariates such as gender, age etc. can respond differently to treatments [8, 39]. This motivates the need for *individualized treatment rules* (ITR) instead of one decision for all. ITR is a deterministic function $\theta(\mathbf{x})$ that maps covariates \mathbf{x} to a treatment t i.e.,

$$\theta(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{T}.$$

Here \mathcal{X} is the set of covariates and \mathcal{T} is the set of treatments. For instance, if $\mathcal{X} = [0, 1]$ and $\mathcal{T} = \{0, 1\}$ then $\theta(x) = \mathbb{1}(x \leq 0.75)$ would map covariates $x \leq 0.75$ to $t = 1$ and covariates $x > 0.75$ to $t = 0$ (see Figure 3.6). The goal is to estimate $\theta(\mathbf{x})$ from observed data $\mathcal{D} = \{(t_i, \mathbf{x}_i, y_i)\}_{i=1}^n$ on treatments t , covariates \mathbf{x} and cost y drawn i.i.d from

$$p(t, \mathbf{x}, y) = p(y|\mathbf{x}, t) \underbrace{p(t|\mathbf{x})}_{\text{past policy}} p(\mathbf{x}). \quad (3.17)$$

In literature, the best ITR is defined as the decision rule which if followed will minimize the average cost [33]. That is,

$$\theta^*(\mathbf{x}) = \arg \min_{\theta(\mathbf{x})} \mathbb{E}^\theta[y]. \quad (3.18)$$

Here \mathbb{E}^θ is expectation with respect to

$$p_\theta(t, \mathbf{x}, y) = p(y|\mathbf{x}, t) \underbrace{\mathbb{1}(t = \theta(\mathbf{x}))}_{\text{new policy}} p(\mathbf{x}). \quad (3.19)$$

Here $\mathbb{1}(t = \theta(\mathbf{x}))$ is a degenerate distribution representing the new policy. Note that we do not observe any data from (3.19) but from (3.17).

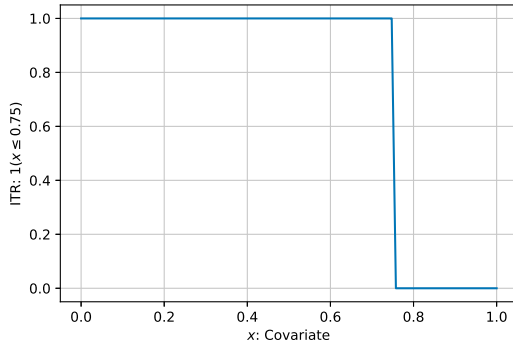


Figure 3.6. An example of individualized treatment rule $\theta(x) = 1(x \leq 0.75)$

However, one can re-write the expectation in (3.18) as

$$\theta^*(\mathbf{x}) = \arg \min_{\theta(\mathbf{x})} \mathbb{E} \left[\frac{\mathbb{1}(t = \theta(\mathbf{x}))}{p(t|\mathbf{x})} y \right], \quad (3.20)$$

where now the expectation is with respect to distribution in (3.17). Moreover, (3.20) can be further simplified into

$$\theta^*(\mathbf{x}) = \arg \min_{t \in \mathcal{T}} \mathbb{E}[y|t, \mathbf{x}]. \quad (3.21)$$

It can be seen that the best ITR for a given covariate \mathbf{x} is the treatment that minimizes the conditional expectation. Hence, one approach used in literature is to learn a model $\hat{\mathbb{E}}[y|t, \mathbf{x}]$ for the conditional expectation and then use that model to find an estimate $\hat{\theta}(\mathbf{x})$ [33]. Let us take a look at an example.

Example 3.3 Here we find an ITR according to (3.21) for the blood pressure data on males and female in Example 2.4. That is,

$$\hat{\theta}(x) = \arg \min_{t \in \{0,1\}} \hat{\mathbb{E}}[y|t, x], \quad (3.22)$$

where $x \in \{0, 1\}$ represents male and female respectively and $\hat{\mathbb{E}}$ denotes the empirical expectation. For $x = 0$, we have

$$\hat{\mathbb{E}}[y|t = 0, x = 0] = 99.9, \quad \hat{\mathbb{E}}[y|t = 1, x = 0] = 105.02 \quad (3.23)$$

and for $x = 1$, we have

$$\hat{\mathbb{E}}[y|t = 0, x = 1] = 99.8, \quad \hat{\mathbb{E}}[y|t = 1, x = 1] = 92.6. \quad (3.24)$$

The ITR $\hat{\theta}(x)$ is

$$\hat{\theta}(x = \text{female}) = 1, \quad \hat{\theta}(x = \text{male}) = 0$$

The success of approaches that specify a model for the conditional expectation of cost y in (3.21) depends on how well-specified the model is. If the model is misspecified then the policy estimate can be biased [33]. Hence, another approach used in literature is to avoid building a model for the conditional expectation. They are based on viewing (3.20) as a weighted classification problem when the treatment is binary i.e.,

$$\theta^*(\mathbf{x}) = \arg \max_{\theta(\mathbf{x})} \mathbb{E} \left[\frac{y}{p(t|\mathbf{x})} \mathbb{1}(t \neq \theta(\mathbf{x})) \right].$$

Here if one thinks of $\theta(\mathbf{x})$ as a classifier then the above equation is equivalent to maximizing a weighted zero-one loss. Several robust approaches have been built along this idea [12, 45, 46].

All the above approaches for learning ITR use the criterion of minimizing the expected cost (3.18). However, in some safety-critical applications, it might be important to minimize the tail cost. For example if y denotes the risk of death of a patient then it would be of interest to have an ITR which minimizes the risk of very sick patients. Using the expected cost criterion in such safety-critical applications might not be appropriate as the following example illustrates.

Example 3.4 *We generate outcome y according to*

$$y = 1 + 3t + x - 5tx + (1 + t + 2tx)u_y, \quad (3.25)$$

where $x \sim \text{Unif}([0, 1])$ is a covariate, t is a binary treatment variable and $\epsilon \sim \mathcal{N}(0, 1)$ is noise. Here y represents negative cost (i.e., reward).

We use the following ITRs to assign treatment based on covariate: (1) $\theta_1(x) = 0$ (2) $\theta_2(x) = 1(x \leq \frac{3}{5})$ (3) $\theta_3(x) = 1(x \leq \frac{1}{2})$ and (4) $\theta_4(x) = 1(x \leq \frac{1}{5})$. The mean optimal ITR in this case is (2). We generate $n = 10^6$ for each ITR and estimate the mean, 25th quantile $Q_{0.25}$ and the 10th quantile $Q_{0.10}$. The results are shown in table 3.1. If we consider a hypothetical setting where y denotes survival time of cancer patients, then the mean optimal ITR (2) may have detrimental effect. ITR (3) gives a higher $Q_{0.25}$ without much degradation in the mean.

To tackle the problem illustrated in Example 3.4, some works target learning ITR that minimize the quantile of the cost directly [20, 43] i.e.,

$$\theta^*(\mathbf{x}) = \arg \min_{\theta(\mathbf{x})} Q_{\tau}^{\theta}(y). \quad (3.26)$$

Here Q_{τ}^{θ} is the τ^{th} quantile of y if treatments were assigned according to ITR $\theta(\mathbf{x})$. These works propose different methods for estimating the above quantile.

ITR	(1)	(2)	(3)	(4)
Mean	1.5	2.39	2.37	2
$Q_{0.25}$	0.8	1.10	1.14	1.01
$Q_{0.10}$	0.16	0.03	0.20	0.33

Table 3.1. Table showing the mean, 25th quantile and the 50th quantile of negative cost i.e., reward for four different policies. The mean optimal policy is (2), however, other policies give higher quantile-rewards without much decrease in the mean reward.

In our work [24], we target minimizing the tail of the cost but we make use of a covariate-treatment specific upper limit on the cost. Specifically, we propose

$$\hat{\theta}(\mathbf{x}) = \arg \min_{t \in \mathcal{T}} y_{\tau}(t, \mathbf{x}). \quad (3.27)$$

Here $y_{\tau}(t, \mathbf{x})$ is an upper limit of the cost y for a given treatment t and covariate \mathbf{x} which satisfies

$$\Pr \{y \leq y_{\tau}(t, \mathbf{x})\} \geq 1 - \tau. \quad (3.28)$$

In words, the idea in (3.27) is that given a person with covariate \mathbf{x} , evaluate the upper limit $y_{\tau}(t, \mathbf{x})$ for each $t \in \mathcal{T}$. Then the optimal treatment is given as the one which gives the smallest $y_{\tau}(t, \mathbf{x})$. Moreover, equation (3.28) indicates that the upper-limit is valid in the sense that the unknown cost lies below $y_{\tau}(t, \mathbf{x})$ with probability atleast $1 - \tau$.

We employ conformal methodology [38, 41] to construct $y_{\tau}(t, \mathbf{x})$ using data $\{(t_i, x_i, y_i)\}_{i=1}^n$ and a simple linear model for y . Unlike [43], our proposed method is applicable even when there are more than two treatments. Moreover, our method also gives a valid upper limit on the cost for the learned ITR.

3.4 Robustness against uncertainty in distribution

So far we have seen examples where the data was drawn from a single data generator. However, in practice one could also obtain data from different *environments* or *contexts*. For instance, data could be collected in different experimental conditions [6], images could be captured in different backgrounds [1] or data could come from different sub-populations [37]. In general,

$$\mathbf{z} \sim p_c(\mathbf{z}),$$

where \mathbf{z} is the data and $p_c(\mathbf{z})$ is the data distribution for context c . When data comes from different contexts, the best decision (2.10) for each context may differ. That is,

$$\theta_c^* = \arg \min_{\theta \in \Theta} \mathbb{E}_{p_c} [\ell_{\theta}(\mathbf{z})]$$

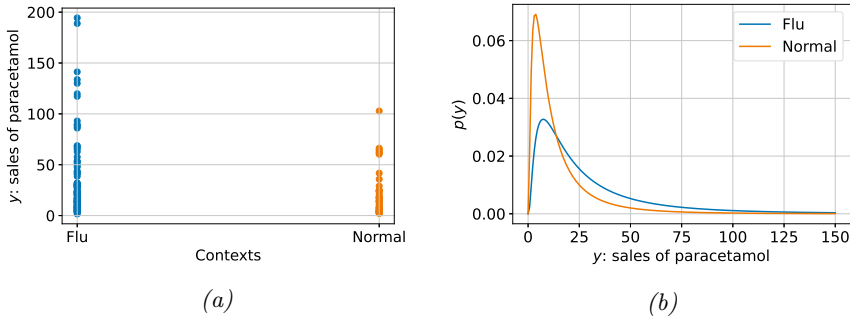


Figure 3.7. Illustration of data from multiple contexts: (a) Scatter plot of data on paracetamol sales y obtained in two different contexts namely when there is a contagious *flu* and when health conditions are *normal* (b) plot of data generating distributions (log-normal) for the two contexts.

where θ_c^* is the best decision and $\mathbb{E}_{p_c}[\ell_{\theta}(z)]$ is the expected loss for context c . Let us take a look at an example.

Example 3.5 Figure 3.7a shows data on sales of paracetamol at a pharmacy obtained in two different contexts namely when there is a contagious flu and when conditions are normal. The data generating process for the two contexts are log-normal distributions with different means shown in Figure 3.7b. The decision variable θ here is how much should the pharmacy stock to minimize their loss. Figure 3.8a shows the expected loss and the corresponding optimal stock levels for each context.

Although each context has its own best decision, however, if the context c is unknown at test time then one needs to make a *single* decision. For instance, the pharmacy might not know whether there will be flu or not next month but it has to decide on how much paracetamol to stock. In such a situation, one approach is that since we do not know what context will occur in the future, let us guard ourselves against the worst that could happen. More precisely,

$$\min_{\theta \in \Theta} \max_{c \in \mathcal{C}} R_c(\theta), \quad (3.29)$$

where

$$R_c(\theta) = \mathbb{E}_{p_c}[\ell_{\theta}(z)],$$

and \mathcal{C} is the set of all possible contexts. This is called the min-max approach [34, 42]. In Figure 3.8a, the best decision for the context *flu* is also the min-max solution.

Another approach for dealing with unknown context is to consider context c to be a random variable and specify a prior distribution over it i.e., $c \sim \pi(c)$. Then minimize the average loss with respect to the prior

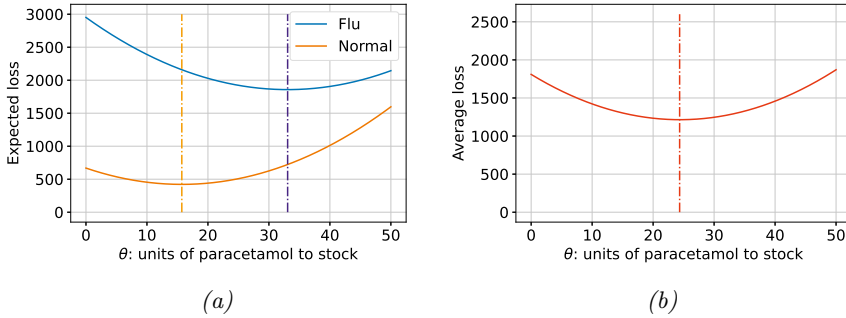


Figure 3.8. (a) Plot of expected loss or risk for the two contexts in the paracetamol data example against units of paracetamol to stock θ . The vertical lines denote the optimal stock level for the two contexts. (b) Plot of average risk with respect to a uniform prior on context c . Vertical line denotes corresponding optimal stock level.

[30] i.e.,

$$\min_{\theta \in \Theta} \sum_{c \in \mathcal{C}} R_c(\theta) \pi(c). \quad (3.30)$$

Figure 3.8b shows the average loss with a uniform prior and the corresponding solution for paracetamol sales data.

Suppose that we have information on context in the training data. For example, for every sales point in Figure 3.7a, we also know whether this was during *flu* period or *normal* period. In general, we observe data point

$$(\mathbf{z}, c) \sim p(\mathbf{z}, c) = p_c(\mathbf{z})p(c),$$

where $p(c)$ is the context distribution.

Figure 3.9a shows the context distribution for paracetamol data. If we assume that at test time, the context distribution is the same as that during training then we can make a single decision according to

$$\begin{aligned} \theta^* &= \arg \min_{\theta \in \Theta} \mathbb{E}_p[\ell_{\theta}(\mathbf{z})], \\ &= \arg \min_{\theta \in \Theta} \sum_{c \in \mathcal{C}} R_c(\theta) p(c), \end{aligned}$$

the ERM version of which is given as

$$\hat{\theta}_{\text{ERM}} = \arg \min_{\theta \in \Theta} \sum_{c \in \mathcal{C}} \hat{R}_c(\theta) \hat{p}(c). \quad (3.31)$$

Here,

$$\hat{p}(c) = \frac{n_c}{n},$$

where n_c is the number of data points observed from context c and n is the total number of data points. Figure 3.9b shows $\hat{\theta}_{\text{ERM}}$ for the data

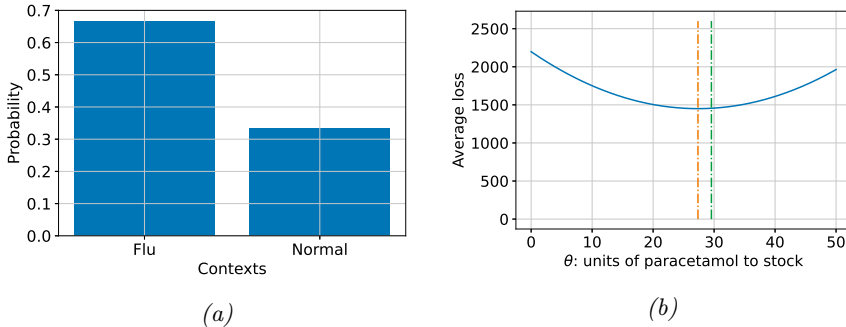


Figure 3.9. (a) Plot of true context distribution $p(c)$ over the two contexts for the paracetamol data in Figure 3.7a. (b) Average risk in (3.31) and two ERM solutions for two different estimates of $\hat{p}(c)$.

on paracetamol sales. The two solutions in Figure 3.9b correspond to two different estimates of $\hat{p}(c)$: one obtained for large n (orange) and the other obtained for small n (green). Note that when n is large, the ERM estimate is close to the minimum. In general, there is *error* in the estimate $\hat{p}(c)$ [5]. When n is large, the error is small and we can go ahead and obtain the ERM estimate in (3.31). However, when n is small the error is larger which leads to poor worst-case performance for ERM.

One approach taken in literature to handle ambiguity of $p(c)$ is to build a distribution set \mathcal{P}_β around $\hat{p}(c)$ as shown in Figure 3.10a. The set is defined as

$$\mathcal{P}_\beta = \{p' : D(\hat{p}, p') \leq \epsilon_\beta\}, \quad (3.32)$$

where $D(\hat{p}, p')$ is some divergence measure between distributions. Many different divergence measures are used in literature such as f-divergences [21], Wasserstein distance [4] etc. Then the idea is to solve the following distribution robust optimization [21] problem

$$\hat{\theta}_{\text{ROB}} = \arg \min_{\theta \in \Theta} \max_{p \in \mathcal{P}_\beta} \sum_{c \in \mathcal{C}} \hat{R}_c(\theta) p(c). \quad (3.33)$$

It is desirable that the set has the property that it covers the true $p(c)$ with high probability i.e.,

$$\Pr\{p(c) \in \mathcal{P}_\beta\} \geq \beta. \quad (3.34)$$

The intuition behind (3.33) is that the set \mathcal{P}_β changes its size depending on the confidence level β and number of data points n (see Figure 3.10b). When n is small, we are uncertain about the estimate of context distribution which is reflected by a large \mathcal{P}_β . Then (3.33) is equivalent to solving a min-max problem. However, when n is large and the estimate of the context distribution is reliable then \mathcal{P}_β is small and concentrated

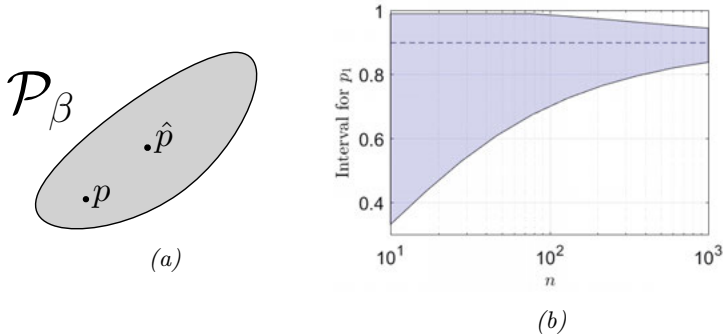


Figure 3.10. Uncertainty in distribution: (a) A schematic of distribution set \mathcal{P}_β around $\hat{p}(c)$. (b) Size of distribution set \mathcal{P}_β against number of data points n for a sample two context problem where $p(c) = [p_1, 1 - p_1] = [0.9, 0.1]$. The vertical axis shows the allowable values of probability of context 1 i.e. p_1 for different values of n .

around the true $p(c)$. In that case, (3.33) is equivalent to (3.31). In this way, one can trade-off between min-max and ERM approaches in a data dependent manner based on the desired level of confidence.

Unlike previous works, which use expected loss or risk $\hat{R}_c(\theta)$ in (3.33), in our work [26] we propose use of excess risk i.e.,

$$\hat{\Delta}_c(\theta) = \hat{R}_c(\theta) - \hat{r}_c,$$

where \hat{r}_c is the minimum risk in context c . This is motivated to counter the pessimistic nature of (3.33) where it could be dominated by a single context which has an overall very high $R_c(\theta)$ [30, 35]. In addition, we construct set \mathcal{P}_β such that it has the coverage property in (3.34) and also adapts size with n as explained above. Sometimes in many of the works in literature, there is no clear criterion for choosing ϵ_β in (3.32).

3.5 Robustness against misspecification

Often in ML, one makes the assumption that the model class is *well-specified*. Well-specified implies that the data generating distribution lies in the model class. For instance, in Example 2.11 where we were fitting a Gaussian distribution, our model class was all Gaussian distributions $\mathcal{N}(z; \mu, \sigma^2)$. And it was well-specified because the temperature was indeed drawn from a Gaussian distribution. However, our model class may very well be *misspecified*. For instance, this would have been the case if the data generating distribution in Example 2.1 would have been a Laplace distribution.

When the model class is misspecified, one needs to be careful about the conclusions drawn from the ERM estimate. The claims that one

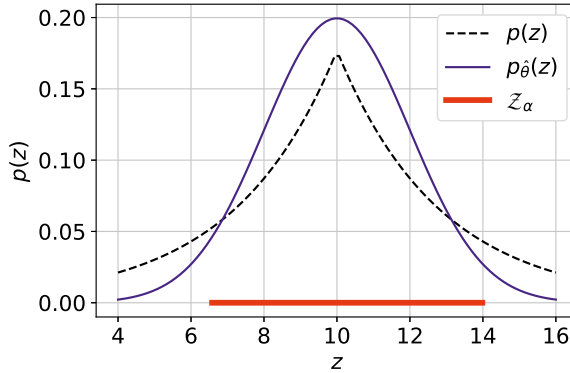


Figure 3.11. An illustration of misspecification in ML. The data generating distribution is Laplace (black dotted) and the assumed model class in Gaussian (solid blue). The prediction interval (red) does not have valid coverage.

makes based on ERM estimates may not be valid. We illustrate this via an example.

Example 3.6 In Example 2.11, we fit a Gaussian distribution to the temperature data z using ERM and the surprisal loss. Now we want to construct a prediction interval \mathcal{Z}_α such that the temperature will lie in it with probability of $1 - \alpha$. That is

$$\Pr\{z \in \mathcal{Z}_\alpha\} = 1 - \alpha. \quad (3.35)$$

We consider two cases:

- *Well-specified:* We draw data from a Gaussian distribution i.e., $p(z) = \mathcal{N}(z; 10, 4)$ and assume a Gaussian model class i.e., $p_\theta(z) = \mathcal{N}(z; \mu, \sigma^2)$.
- *Misspecified:* We draw data from a Laplace distribution i.e., $p(z) = \text{Laplace}(z; 10, \sqrt{2} * 2)$ and assume a Gaussian model class.

We estimate $\hat{\mu}$ and $\hat{\sigma}$ using ERM and construct the prediction interval as

$$\mathcal{Z}_\alpha = [\hat{\mu} - q_{\alpha/2}\hat{\sigma}, \hat{\mu} + q_{\alpha/2}\hat{\sigma}], \quad (3.36)$$

where $q_{\alpha/2} = Q^{-1}(\frac{\alpha}{2})$ is the inverse q -function. We draw a test point z from Gaussian and Laplace distributions and check if it lies in the prediction interval. We compute the coverage probability in (3.35) by repeating the process over many Monte carlo simulations. We choose $\alpha = 1\%$. The coverage for the two cases turns out to be

$$\widehat{\Pr}\{z \in \mathcal{Z}_\alpha\} = \begin{cases} 99.7\% & \text{Well-specified,} \\ 96.7\% & \text{Misspecified.} \end{cases} \quad (3.37)$$

It can be seen that the coverage in the case of misspecified model class is not what we expect it to be. In other words, the prediction interval is not valid.

Conformal prediction is a method for constructing *valid* prediction intervals [41] \mathcal{Z}_α . The conformal method uses a *predictor* for the quantity of interest and *scores* to determine whether a new data point *conforms* with the observed data. For instance, in Example 3.6, the predictor for temperature could be the sample mean

$$\bar{z}_n = \frac{1}{n} \sum_{i=1}^n z_i \quad (3.38)$$

where $\mathcal{D} = \{z_i\}_{i=1}^n$ is the data. In conformal method, a candidate point z_{n+1} is augmented to the data to create an augmented data set of $n + 1$ points. This gives us a corresponding updated sample mean \bar{z}_{n+1} . Then scores are computed for all the data points i.e.,

$$s_i = |z_i - \bar{z}_{n+1}| \text{ for } i = 1, \dots, n + 1. \quad (3.39)$$

These scores are used to find a rank for the candidate point i.e.,

$$r(z_{n+1}) = \frac{1}{n + 1} \sum_{i=1}^{n+1} \mathbb{1}(s_i \leq s_{n+1}). \quad (3.40)$$

The rank $r(z_{n+1})$ and the coverage level α are used to determine whether or not the data point z_{n+1} is part of the interval or not. The entire prediction interval is built by repeatedly checking for new candidate points in some range. The conformal prediction interval for the misspecified case in Example 3.6, gives a coverage of 99.6% for $\alpha = 1\%$.

In our work [25], we make use of the conformal methodology to build valid prediction intervals for predicting count data across space (e.g. number of crimes in a district). We also propose a regularized maximum likelihood approach which, empirically, gives smaller interval size as compared to the unregularized counterpart.

References

- [1] M. Arjovsky, L. Bottou, et al. “Invariant risk minimization”. In: *arXiv preprint arXiv:1907.02893* (2019).
- [2] J. Besag, J. York, et al. “Bayesian image restoration, with two applications in spatial statistics”. In: *Annals of the institute of statistical mathematics* 43.1 (1991), pp. 1–20.
- [3] C. M. Bishop. “Pattern Recognition and Machine Learning”. Springer, 2006.
- [4] J. Blanchet, Y. Kang, et al. “Robust Wasserstein profile inference and applications to machine learning”. In: *Journal of Applied Probability* 56.3 (2019), pp. 830–857.
- [5] R. Bradley. “Decision theory with a human face”. Cambridge University Press, 2017.
- [6] P. Bühlmann. “Invariance, Causality and Robustness”. In: *Statistical Science* 35.3 (2020), pp. 404–426.
- [7] H. Cantzler. “Random sample consensus (RANSAC)”. In: *Institute for Perception, Action and Behaviour, Division of Informatics, University of Edinburgh* (1981).
- [8] B. Chakraborty and E. Moodie. “Statistical methods for dynamic treatment regimes”. In: *Springer-Verlag. doi 10* (2013), pp. 978–1.
- [9] T. M. Cover. “Elements of information theory”. John Wiley & Sons, 1999.
- [10] N. Cressie and C. K. Wikle. “Statistics for spatio-temporal data”. John Wiley & Sons, 2011.
- [11] I. Diakonikolas, G. Kamath, et al. “Sever: A robust meta-algorithm for stochastic optimization”. In: *International Conference on Machine Learning*. PMLR, 2019, pp. 1596–1606.
- [12] S. Fu, Q. He, et al. “Robust outcome weighted learning for optimal individualized treatment rules”. In: *Journal of biopharmaceutical statistics* 29.4 (2019), pp. 606–624.
- [13] M. Glymour, J. Pearl, et al. “Causal inference in statistics: A primer”. John Wiley & Sons, 2016.

- [14] E. M. Hanks, E. M. Schliep, et al. “Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification”. In: *Environmetrics* 26.4 (2015), pp. 243–254.
- [15] R. J. Hernán MA. “Causal Inference: What If.” Chapman & Hall/CRC, 2020.
- [16] J. S. Hodges and B. J. Reich. “Adding spatially-correlated errors can mess up the fixed effect you love”. In: *The American Statistician* 64.4 (2010), pp. 325–334.
- [17] P. J. Huber and E. M. Ronchetti. “Robust Statistics”. Wiley, 2009.
- [18] J. Hughes and M. Haran. “Dimension reduction and alleviation of confounding for spatial generalized linear mixed models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.1 (2013), pp. 139–159.
- [19] K. A. Lai, A. B. Rao, et al. “Agnostic estimation of mean and covariance”. In: IEEE. 2016, pp. 665–674.
- [20] K. A. Linn, E. B. Laber, et al. “Interactive q-learning for quantiles”. In: *Journal of the American Statistical Association* 112.518 (2017), pp. 638–649.
- [21] H. Namkoong and J. C. Duchi. “Stochastic gradient methods for distributionally robust optimization with f-divergences”. In: *Advances in neural information processing systems*. 2016, pp. 2208–2216.
- [22] J. Nocedal and S. J. Wright. “Numerical optimization”. Springer, 1999.
- [23] M. Osama, D. Zachariah, et al. “Inferring heterogeneous causal effects in presence of spatial confounding”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4942–4950.
- [24] M. Osama, D. Zachariah, et al. “Learning robust decision policies from observational data”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 18205–18214.
- [25] M. Osama, D. Zachariah, et al. “Prediction of spatial point processes: regularized method with out-of-sample guarantees”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [26] M. Osama, D. Zachariah, et al. “Robust learning in heterogeneous contexts”. In: *arXiv preprint arXiv:2105.08532* (2022).
- [27] M. Osama, D. Zachariah, et al. “Robust localization in wireless networks from corrupted signals”. In: *EURASIP Journal on Advances in Signal Processing* 2021.1 (2021), pp. 1–18.

- [28] M. Osama, D. Zachariah, et al. “Robust risk minimization for statistical learning from corrupted data”. In: *IEEE Open Journal of Signal Processing* (2020), pp. 287–294.
- [29] C. J. Paciorek. “The importance of scale for spatial-confounding bias and precision of spatial regression estimators”. In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 25.1 (2010), p. 107.
- [30] G. Parmigiani and L. Inoue. “Decision theory: Principles and approaches”. John Wiley & Sons, 2009.
- [31] J. Peters, D. Janzing, et al. “Elements of causal inference: foundations and learning algorithms”. The MIT Press, 2017.
- [32] A. Prasad, A. S. Suggala, et al. “Robust estimation via robust gradient estimation”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2020), pp. 601–627.
- [33] M. Qian and S. A. Murphy. “Performance guarantees for individualized treatment rules”. In: *Annals of statistics* 39.2 (2011), p. 1180.
- [34] S. Sagawa, P. W. Koh, et al. “Distributionally Robust Neural Networks”. In: *International Conference on Learning Representations*. 2020.
- [35] L. J. Savage. “The theory of statistical decision”. In: *Journal of the American Statistical association* 46.253 (1951), pp. 55–67.
- [36] A. Shapiro, D. Dentcheva, et al. “Lectures on stochastic programming: modeling and theory”. SIAM, 2021.
- [37] H. Shimodaira. “Improving predictive inference under covariate shift by weighting the log-likelihood function”. In: *Journal of statistical planning and inference* 90.2 (2000), pp. 227–244.
- [38] R. J. Tibshirani, R. Foygel Barber, et al. “Conformal prediction under covariate shift”. In: *Advances in neural information processing systems* 32 (2019).
- [39] A. A. Tsiatis, M. Davidian, et al. “Dynamic treatment regimes: Statistical methods for precision medicine”. Chapman and Hall/CRC, 2019.
- [40] V. N. Vapnik. “Statistical Learning Theory”. Wiley, 1998.
- [41] V. Vovk, A. Gammerman, et al. “Algorithmic learning in a random world”. Springer Science & Business Media, 2005.
- [42] A. Wald. “Statistical decision functions. Wiley”. In: (1950).
- [43] L. Wang, Y. Zhou, et al. “Quantile-optimal treatment regimes”. In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1243–1254.

- [44] L. Wasserman. “All of Statistics: A concise course in Statistical Inference”. Springer, 2005.
- [45] Y. Zhao, D. Zeng, et al. “Estimating individualized treatment rules using outcome weighted learning”. In: *Journal of the American Statistical Association* 107.499 (2012), pp. 1106–1118.
- [46] X. Zhou, N. Mayer-Hamblett, et al. “Residual weighted learning for estimating individualized treatment rules”. In: *Journal of the American Statistical Association* 112.517 (2017), pp. 169–187.
- [47] A. M. Zoubir, V. Koivunen, et al. “Robust statistics for signal processing”. Cambridge University Press, 2018.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2147*

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-472453



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2022