

Machine Learning and Neural Networks III (MATH3431)

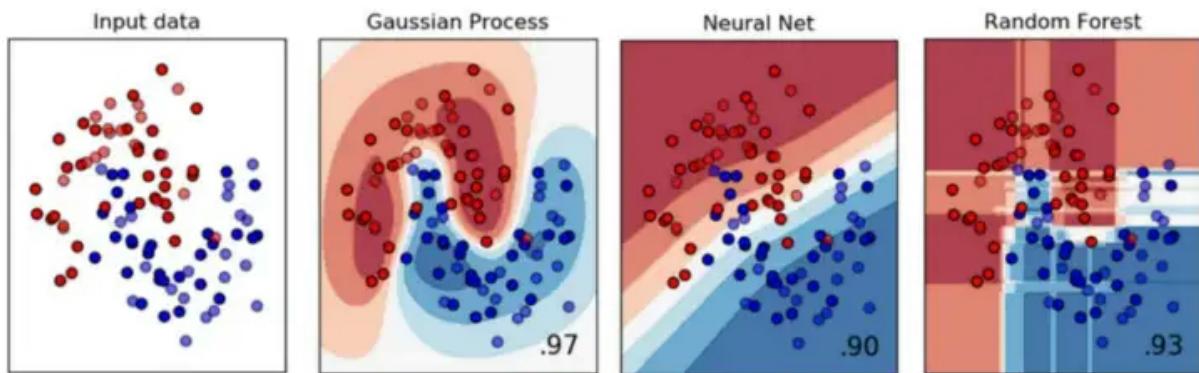
Epiphany term

Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Department of Mathematical Sciences (Office MCS3088)
Durham University
Stockton Road Durham DH1 3LE UK

2023/03/14 at 17:47:51



Reading list

These lecture Handouts have been derived based on the above reading list.

Main texts:

- Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.
 - It is a classical textbook in machine learning (ML) methods. It discusses all the concepts introduced in the course (not necessarily in the same depth). It is one of the main textbooks in the module. The level on difficulty is easy.
 - Students who wish to have a textbook covering traditional concepts in machine learning are suggested to get a copy of this textbook. It is available online from the Microsoft's website <https://www.microsoft.com/en-us/research/publication/pattern-recognition-and-machine-learning-by-cristian-m-bishop/>
- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
 - It has several elements of theory about machine learning algorithms. It is one of the main textbooks in the module. The level on difficulty is advanced as it requires moderate knowledge of maths.
- Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.
 - It is a classical textbook about ‘traditional’ artificial neural networks (ANN). It is very comprehensive (compared to others) and it goes deep enough for the module although it may be a bit outdated. It is one of the main textbooks in the module for ANN. The level on difficulty is moderate.

Supplementary textbooks:

- Ripley, B. D. (2007). Pattern recognition and neural networks. Cambridge university press.
 - A classical textbook in artificial neural networks (ANN) that also covers other machine learning concepts. It contains interesting theory about ANN.
 - It is suggested to be used as a supplementary reading for neural networks as it contains a few interesting theoretical results. The level on difficulty is moderate.
- Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.
 - A classic book in Gaussian process regression (GPR) that covers the material we will discuss in the course about GPR. It can be used as a companion textbook with that of (Bishop, C. M., 2006). The level on difficulty is easy.

- Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.
 - A popular textbook in machine learning methods. It discusses all the concepts introduced in the module. It focuses more on the probabilistic/Bayesian framework but not with great detail. It can be used as a comparison textbook for brief reading about ML methods just to see another perspective than that in (Bishop, C. M., 2006). The level on difficulty is easy.
- Murphy, K. P. (2022). Probabilistic machine learning: an introduction. MIT press.
 - A textbook in machine learning methods. It covers a smaller number of ML concepts than (Murphy, K. P., 2012) but it contains more fancy/popular topics such as deep learning ideas. It is suggested to be used in the same manner as (Murphy, K. P., 2012). The level on difficulty is easy.
- Barber, D. (2012). Bayesian reasoning and machine learning. Cambridge University Press.
 - A textbook in machine learning methods from a Bayesian point of view. It discusses all the concepts introduced apart from ANN and stochastic gradient algorithms. It aims to be more ‘statistical’ than those of Murphy and Bishop. The level on difficulty is easy.
- Devroye, L., Györfi, L., & Lugosi, G. (2013). A probabilistic theory of pattern recognition (Vol. 31). Springer Science & Business Media.
 - Theoretical aspects about machine learning algorithms. The level on difficulty is advanced as it requires moderate knowledge of probability.

Handout 0: Learning problem: Definitions, notation, and formulation –A recap

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To get some definitions and set-up about the learning procedure; essentially to formalize what introduced in term 1.

Reading list & references:

- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
– Ch. 1 Introduction
- Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.
– Ch. 1 Introduction

1. GENERAL INTRODUCTIONS AND LOOSE DEFINITIONS

Pattern recognition is the automated discovery of patterns and regularities in data $z \in \mathcal{Z}$. **Machine learning (ML)** are statistical procedures for building and understanding probabilistic methods that 'learn'. **ML algorithms** build a (probabilistic/deterministic) model able to make predictions or decisions with minimum human interference and can be used for pattern recognition. **Learning** (or training, estimation) is called the procedure where the ML model is tuned. **Training data** (or observations, sample data set, examples) is a set of observables $\{z_i \in \mathcal{Z}\}$ used to tune the parameters of the ML model. By \mathcal{Z} we denote the examples (or observables) domain. **Test set** is a set of available examples/observables $\{z'_i\}$ (different than the training data) used to verify the performance of the ML model for a given a measure of success. **Measure of success** (or performance) is a quantity that indicates how bad the corresponding ML model or Algorithm performs (eg quantifies the failure/error), and can also be used for comparisons among different ML models; eg, **Risk function** or **Empirical Risk Function**. Two main problems in ML are the supervised learning (we focus here) and the unsupervised learning.

Supervised learning problems involve applications where the training data $z \in \mathcal{Z}$ comprises examples of the input vectors $x \in \mathcal{X}$ along with their corresponding target vectors $y \in \mathcal{Y}$; i.e. $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. By \mathcal{X} we denote the inputs (or instances) domain, and by \mathcal{Y} we denote the target domain. **Classification problems** are those which aim to assign each input vector x to one of a finite number of discrete categories of y . **Regression problems** are those where the output y consists of one or more continuous variables. All in all, the learner wishes to recover an unknown pattern (i.e. functional relationship) between components $x \in \mathcal{X}$ that serves as inputs and components $y \in \mathcal{Y}$ that act as outputs; i.e. $x \mapsto y$. Hence, \mathcal{X} is the input domain, and \mathcal{Y} is the output (or target) domain. The goal of learning is to discover a function which predicts $y \in \mathcal{Y}$ from $x \in \mathcal{X}$.

Unsupervised learning problems involve applications where the training data $z \in \mathcal{Z}$ consist of a set of input vectors $x \in \mathcal{X}$ without any corresponding target values ; i.e. $\mathcal{Z} = \mathcal{X}$. In clustering the goal is to discover groups of similar examples within the data of it is to discover groups of similar examples within the data.

2. (LOOSE) NOTATION & DEFINITIONS IN LEARNING

Definition 1. The learner's output is a function, $h : \mathcal{X} \rightarrow \mathcal{Y}$ which predicts $y \in \mathcal{Y}$ from $x \in \mathcal{X}$. It is also called hypothesis, prediction rule, predictor, or classifier.

Notation 2. We often denote the set of hypothesis as \mathcal{H} ; i.e. $h \in \mathcal{H}$.

Example 3. (Linear Regression)¹ Consider the regression problem where the goal is to learn the mapping $x \rightarrow y$ where $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$. Hypothesis is a linear function $h : \mathcal{X} \rightarrow \mathcal{Y}$ (that learner wishes to learn) to approximate mapping $x \rightarrow y$. The hypothesis set $\mathcal{H} = \{x \rightarrow \langle w, x \rangle : w \in \mathbb{R}^d\}$. We can use the loss $\ell(h, (x, y)) = (h(x) - y)^2$.

Definition 4. Training data set \mathcal{S} of size m is any finite sequence of pairs $((x_i, y_i); i = 1, \dots, m)$ in $\mathcal{X} \times \mathcal{Y}$. This is the information that the learner has assess.

Definition 5. Data generation model $g(\cdot)$ is the probability distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, unknown to the learner that has generated the data.

Definition 6. We denote as $\mathfrak{A}(\mathcal{S})$ the hypothesis (outcome) that a learning algorithm \mathfrak{A} returns given training sample S .

Definition 7. (Loss function) Given any set of hypothesis \mathcal{H} and some domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, a loss function $\ell(\cdot)$ is any function $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}_+$. The purpose of loss function $\ell(h, z)$ is to quantify the “error” for a given hypothesis h and example z –the greater the error the greater its value of the loss.

Example 8. (Cont. Example 3) In regression problems $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y} \subset \mathbb{R}$ is uncountable, a loss function can be

$$\ell_{\text{sq}}(h, (x, y)) = (h(x) - y)^2$$

Example 9. In binary classification problems with $h : \mathcal{X} \rightarrow \mathcal{Y}$ a learner where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y} = \{0, 1\}$ is discrete, a loss function can be

$$\ell_{0-1}(h, (x, y)) = 1(h(x) \neq y),$$

Definition 10. (Risk function) The risk function $R_g(h)$ of h is the expected loss of the hypothesis $h \in \mathcal{H}$, w.r.t. probability distribution g over domain Z ; i.e.

$$(2.1) \quad R_g(h) = \mathbb{E}_{z \sim g}(\ell(h, z))$$

¹ $\langle w, x \rangle = w^\top x$

Remark 11. In learning, an ideal way to obtain an optimal predictor h^* is to compute the risk minimizer

$$h^* = \arg \min_{\forall h} (R_g(h))$$

Example 12. (Cont. Ex. 8) The risk function is $R_g(h) = E_{z \sim g}(h(x) - y)^2$, and it measures the quality of the hypothesis function $h : \mathcal{X} \rightarrow \mathcal{Y}$, (or equiv. the validity of the class of hypotheses \mathcal{H}) against the data generating model g , as the expected square difference between the predicted values from h and the true target values y at every x .

Note 13. Computing the risk minimizer may be practically challenging due to the integration w.r.t. the unknown data generation model g involved in the expectation (2.1). Sub-optimally, one may resort to the Empirical risk function.

Definition 14. (Empirical risk function) The empirical risk function $\hat{R}_S(h)$ of h is the expectation of loss of h over a given sample $S = (z_1, \dots, z_m) \in \mathcal{Z}^m$; i.e.

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, z_i).$$

Example 15. (Cont. Example 12) Given given sample $S = \{(x_i, y_i); i = 1, \dots, m\}$ the empirical risk function is $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2$.

Example 16. Consider a learning problem where the true data generation distribution (unknown to the learner) is $g(z)$, the statistical model (known to the learner) is given by a sampling distribution $f_\theta(y) := f(y|\theta)$ labeled by an unknown parameter θ . The goal is to learn θ . If we assume loss function

$$\ell(\theta, z) = \log \left(\frac{g(z)}{f_\theta(z)} \right)$$

then the risk is

$$(2.2) \quad R_g(\theta) = E_{z \sim g} \left(\log \left(\frac{g(z)}{f_\theta(z)} \right) \right) = E_{z \sim g} (\log(g(z))) - E_{z \sim g} (\log(f_\theta(z)))$$

whose minimizer is

$$\theta^* = \arg \min_{\forall \theta} (R_g(\theta)) = \arg \min_{\forall \theta} (E_{z \sim g} (-\log(f_\theta(z))))$$

as the first term in (2.2) is constant. Note that in the Maximum Likelihood Estimation technique the MLE θ_{MLE} is the minimizer of

$$\theta_{MLE} = \arg \min_{\forall \theta} \left(\frac{1}{m} \sum_{i=1}^m (-\log(f_\theta(z_i))) \right)$$

where $S = \{z_1, \dots, z_m\}$ is an IID sample from g . Hence, MLE θ_{MLE} can be considered as the minimizer of the empirical risk $R_S(\theta) = \frac{1}{m} \sum_{i=1}^m (-\log(f_\theta(z_i)))$.

Definition 17. A learning problem with hypothesis class \mathcal{H} , examples domain \mathcal{Z} , and loss function ℓ may be denoted with a triplet $(\mathcal{H}, \mathcal{Z}, \ell)$.

Example 18. Consider the multiple linear regression problem $x \mapsto y$ where $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$. Till now, we set the learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ in the linear regression with hypothesis class $\mathcal{H} = \{x \rightarrow \langle w, x \rangle : w \in \mathbb{R}^d\}$, loss $\ell(h, (x, y)) = (h(x) - y)^2$, and examples domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. Because learning problem involves only linear functions as predictors $h(x) = \langle w, x \rangle$, this learning problem could be defined equivalently with a hypothesis class $\mathcal{H} = \{w \in \mathbb{R}^d\}$ and loss function loss $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$. The latter will be mainly used.

APPENDIX A. USEFUL BITS

Below are some standard notation used as default in the notes except in cases that is defined otherwise.

- q -norm: When $x \in \mathbb{R}^d$ $\|x\|_q := \left(\sum_{j=1}^d x_j^q \right)^{1/q}$
- Manhattan norm: When $x \in \mathbb{R}^d$ $\|x\|_1 := \sum_{j=1}^d |x_j|$
- Euclidean norm: When $x \in \mathbb{R}^d$ $\|x\|_2 := \sqrt{\sum_{j=1}^d x_j^2}$. When $\|\cdot\|$ we will assume the Euclidean norm.
- Infinity norm or maximum norm: $\|x\|_\infty := \max_{\forall j} |x_j|$
- Inner product of x, y : If $x, y \in \mathbb{R}^d$ then $\langle x, y \rangle = x^\top y$. So $\langle x, x \rangle = \|x\|^2$

Also some standard formulas.

- Jensens' inequality: If $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ then

$$\begin{cases} f(\mathbf{E}(x)) \leq \mathbf{E}(f(x)) & \text{if } f \text{ is convex} \\ f(\mathbf{E}(x)) \geq \mathbf{E}(f(x)) & \text{if } f \text{ is concave} \end{cases}$$

- Cauchy–Schwarz inequality: If $x, y \in \mathbb{R}^d$ then $|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$ equiv. $|\langle x, y \rangle| \leq \|x\| \|y\|$.

Handout 1: Elements of convex learning problems

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce elements of convexity, Lipschitzness, and smoothness that can be used for the analysis of stochastic gradient related learning algorithms.

Reading list & references:

- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
 - Ch. 12 Convex Learning Problems

Further reading

- Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.

1. MOTIVATIONS

Note 1. We introduce concepts of convexity and smoothness that facilitate the (theoretical) analysis of the learning problems and their solution that we will discuss (eg stochastic gradient descent) later on. Also learning problems with such characteristics can be learned more efficiently.

Note 2. Most of the ML problems discussed in the course (eg, Artificial neural networks, Gaussian process regression) are usually non-convex.

Note 3. To overcome this problem, we will introduce the concept of surrogate loss function that allows a non-convex problem to be handled with the tools introduced in the convex setting.

2. CONVEX LEARNING PROBLEMS

Note 4. Convex learning problem is a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ that the hypothesis class \mathcal{H} is a convex set, and the loss function ℓ is a convex function for each example $z \in \mathcal{Z}$.

Example 5. Multiple linear regression $\langle w, x \rangle \rightarrow y$ with $y \in \mathbb{R}$, hypothesis class $\mathcal{H} = \{w \in \mathbb{R}^d\}$ and loss $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$ with

$$w^* = \arg \min_{\forall w} E(\langle w, x \rangle - y)^2$$

or

$$w^{**} = \arg \min_{\forall w} \frac{1}{m} \sum_{i=1}^m (\langle w, x_i \rangle - y)^2$$

is a convex learning problem for reasons that will be discussed below.

Moved
to
Section 6
Moved
to
Section 6
Moved
to
Section 6

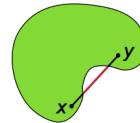
3. CONVEXITY

Definition 6. A set C is convex if for any $u, v \in C$, the line segment between u and v is contained in C . Namely,

- for any $u, v \in C$ and for any $\alpha \in [0, 1]$ we have that $\alpha u + (1 - \alpha) v \in C$.



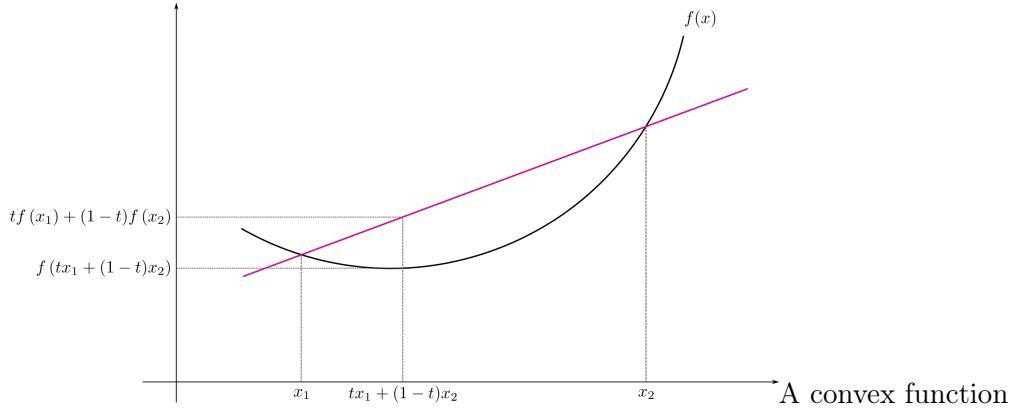
A convex set



A non-convex set

Definition 7. Let C be a convex set. A function $f : C \rightarrow \mathbb{R}$ is convex function if for any $u, v \in C$ and for any $\alpha \in [0, 1]$

$$f(\alpha u + (1 - \alpha) v) \leq \alpha f(u) + (1 - \alpha) f(v)$$



Example 8. The function $f : \mathbb{R} \rightarrow \mathbb{R}_+$ with $f(x) = x^2$ is convex function. For any $u, v \in C$ and for any $\alpha \in [0, 1]$ it is

$$(\alpha u + (1 - \alpha) v)^2 - \alpha(u)^2 + (1 - \alpha)(v)^2 = -\alpha(1 - \alpha)(u - v)^2 \leq 0$$

Proposition 9. Every local minimum of a convex function is the global minimum.

Proposition 10. Let $f : C \rightarrow \mathbb{R}$ be convex function. The tangent of f at $w \in C$ is below f , namely

$$\forall u \in C \quad f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle$$

Proposition 11. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $f(w) = g(\langle w, x \rangle + y)$ for some $x \in \mathbb{R}^d$, $y \in \mathbb{R}$. If g is convex function then f is convex function.

Proof. See Exercise 1 in the Exercise sheet. □

Example 12. Consider the regression problem $x \mapsto y$ with $x \in \mathbb{R}^d$, $y \in \mathbb{R}$ and predictor $h(x) = \langle w, x \rangle$. The loss function $\ell(w, (x, y)) = (\langle w, x \rangle + y)^2$ is convex because $g(a) = (a)^2$ is convex and Proposition 11.

Example 13. Let $f_j : \mathbb{R}^d \rightarrow \mathbb{R}$ convex functions for $j = 1, \dots, r$. Then:

- (1) $g(x) = \max_{\forall j} (f_j(x))$ is a convex function

(2) $g(x) = \sum_{j=1}^r w_j f_j(x)$ is a convex function where $w_j > 0$

Solution.

(1) For any $u, v \in \mathbb{R}^d$ and for any $\alpha \in [0, 1]$

$$\begin{aligned} g(\alpha u + (1 - \alpha)v) &= \max_{\forall j} (f_j(\alpha u + (1 - \alpha)v)) \\ &\leq \max_{\forall j} (\alpha f_j(u) + (1 - \alpha)f_j(v)) && (f_j \text{ is convex}) \\ &\leq \alpha \max_{\forall j} (f_j(u)) + (1 - \alpha) \max_{\forall j} (f_j(v)) && (\max(\cdot) \text{ is convex}) \\ &\leq \alpha g(u) + (1 - \alpha)g(v) \end{aligned}$$

(2) For any $u, v \in \mathbb{R}^d$ and for any $\alpha \in [0, 1]$

$$\begin{aligned} g(\alpha u + (1 - \alpha)v) &= \sum_{j=1}^r w_j f_j(\alpha u + (1 - \alpha)v) \\ &\leq \alpha \sum_{j=1}^r w_j f_j(u) + (1 - \alpha) \sum_{j=1}^r w_j f_j(v) && (f_j \text{ is convex}) \\ &\leq \alpha g(u) + (1 - \alpha)g(v) \end{aligned}$$

Example 14. $g(x) = |x|$ is convex according to Example 13, as $g(x) = |x| = \max(-x, x)$.

4. LIPSCHITZBNESS

Definition 15. Let $C \in \mathbb{R}^d$. Function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is ρ -Lipschitz over C if for every $w_1, w_2 \in C$ we have that

$$(4.1) \quad \|f(w_1) - f(w_2)\| \leq \rho \|w_1 - w_2\|. \quad \text{Lipschitz condition}$$

Conclusion 16. That means: a Lipschitz function $f(x)$ cannot change too drastically wrt x .

Example 17. Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}_+$ with $f(x) = x^2$.

- (1) f is not a ρ -Lipschitz in \mathbb{R} .
- (2) f is a ρ -Lipschitz in $C = \{x \in \mathbb{R} : |x| < \rho/2\}$.

$$|f(x_2) - f(x_1)| = |x_2^2 - x_1^2| = |(x_2 + x_1)(x_2 - x_1)| \leq 2\rho/2(x_2 - x_1) = \rho|x_2 - x_1|$$

Solution.

- (1) For $x_1 = 0$ and $x_2 = 1 + \rho$, it is

$$|f(x_2) - f(x_1)| = (1 + \rho)^2 > \rho(1 + \rho) = \rho|x_2 - x_1|$$

- (2) It is

$$|f(x_2) - f(x_1)| = |x_2^2 - x_1^2| = |(x_2 + x_1)(x_2 - x_1)| \leq 2\rho/2(x_2 - x_1) = \rho|x_2 - x_1|$$

Theorem 18. Let functions g_1 be ρ_1 -Lipschitz and g_2 be ρ_2 -Lipschitz. Then f with $f(x) = g_1(g_2(x))$ is $\rho_1\rho_2$ -Lipschitz.

Solution. See Exercise 2 from the exercise sheet

Example 19. Let functions g be ρ -Lipschitz. Then f with $f(x) = g(\langle v, x \rangle + b)$ is $(\rho|v|)$ -Lipschitz.

$$\begin{aligned} |f(w_1) - f(w_2)| &= |g(\langle v, w_1 \rangle + b) - g(\langle v, w_2 \rangle + b)| \leq \rho |\langle v, w_1 \rangle + b - \langle v, w_2 \rangle - b| \\ &\leq \rho |v^\top w_1 - v^\top w_2| \leq \rho |v| |w_1 - w_2| \end{aligned}$$

Note 20. So, given Examples 17 and 19, in the linear regression setting using loss $\ell(w, z) = (w^\top x - z)^2$, the loss function is β -Lipschitz for a given z and bounded $\|w\| < \rho$.

5. SMOOTHNESS

Definition 21. A differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is β -smooth if its gradient is β -Lipschitz; namely for all $v, w \in \mathbb{R}^d$

$$(5.1) \quad \|\nabla f(w_1) - \nabla f(w_2)\| \leq \beta \|w_1 - w_2\|.$$

Theorem 22. Function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is β -smooth iff

$$(5.2) \quad f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2$$

Remark 23. If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is β -smooth then (5.2) holds, and if it is convex as well then

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle$$

holds. Hence if both conditions imply upper and lower bounds

$$f(v) - f(w) \in \left(\langle \nabla f(w), v - w \rangle, \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2 \right)$$

Remark 24. If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is β -smooth then for $v, w \in \mathbb{R}^d$ such that $v = w - \frac{1}{\beta} \nabla f(w)$ then by (5.2), it is

$$\frac{1}{2\beta} \|\nabla f(w)\|^2 \leq f(w) - f(v)$$

If additionally $f(x) > 0$ for all $x \in \mathbb{R}^d$ then

$$\|\nabla f(w)\|^2 \leq 2\beta f(w)$$

which provides assumptions to bound the gradient.

Theorem 25. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with $f(w) = g(\langle w, x \rangle + y)$ $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a β -smooth function. Then f is a $(\beta \|x\|^2)$ -smooth.

Solution. See Exercise 3 from the Exercise sheet

Example 26. Let $f(w) = (\langle w, x \rangle + y)^2$ for $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Then f is $(2\|x\|^2)$ -smooth.

Solution. It is $f(w) = g(\langle w, x \rangle + y)$ for $g(a) = a^2$. g is 2-smooth since

$$\|g'(w_1) - g'(w_2)\| = \|2w_1 - 2w_2\| \leq 2 \|w_1 - w_2\|.$$

Hence from (25), f is $\left(2\|x\|^2\right)$ -smooth.

6. CONVEX LEARNING PROBLEMS

Definition 27. Convex learning problem is a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ that the hypothesis class \mathcal{H} is a convex set, and the loss function ℓ is a convex function for each example $z \in \mathcal{Z}$.

Definition 28. Convex-Lipschitz-Bounded Learning Problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with parameters ρ , and B , is called the learning problem whose the hypothesis class \mathcal{H} is a convex set, for all $w \in \mathcal{H}$ it is $\|w\| \leq B$, and the loss function $\ell(\cdot, z)$ is convex and ρ -Lischitz function for all $z \in \mathcal{Z}$.

Definition 29. Convex-Smooth-Bounded Learning Problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with parameters β , and B , is called the learning problem whose the hypothesis class \mathcal{H} is a convex set, for all $w \in \mathcal{H}$ it is $\|w\| \leq B$, and the loss function $\ell(\cdot, z)$ is convex, nonnegative, and β -smooth function for all $z \in \mathcal{Z}$.

Example 30. Multiple linear regression $\langle w, x \rangle \rightarrow y$ with $y \in \mathbb{R}$, hypothesis class $\mathcal{H} = \{w \in \mathbb{R}^d\}$ and loss $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$ with

$$w^* = \arg \min_{\forall w} \mathbb{E}(\langle w, x \rangle - y)^2$$

or

$$w^{**} = \arg \min_{\forall w} \frac{1}{m} \sum_{i=1}^m (\langle w, x_i \rangle - y)^2$$

is a convex learning problem.

7. NON-CONVEX LEARNING PROBLEMS (SURROGATE TREATMENT)

Remark 31. A learning problem may involve non-convex loss function $\ell(w, z)$ which implies a non-convex risk function $R_g(w)$. However, our learning algorithm will be analyzed in the convex setting. A suitable treatment to overcome this difficulty would be to upper bound the non-convex loss function $\ell(w, z)$ by a convex surrogate loss function $\tilde{\ell}(w, z)$ for all w , and use $\tilde{\ell}(w, z)$ instead of $\ell(w, z)$.

Example 32. Consider the binary classification problem with inputs $x \in \mathcal{X}$, outputs $y \in \{-1, +1\}$; we need to learn $w \in \mathcal{H}$ from hypothesis class $\mathcal{H} \subset \mathbb{R}^d$ with respect to the loss

$$\ell(w, (x, y)) = 1_{(y\langle w, x \rangle \leq 0)}$$

with $y \in \mathbb{R}$, and $x \in \mathbb{R}^d$. Here $\ell(\cdot)$ is non-convex. A convex surrogate loss function can be

$$\tilde{\ell}(w, (x, y)) = \max(0, 1 - y\langle w, x \rangle)$$

which is convex (Example 14) wrt w . Note that:

- $\tilde{\ell}(w, (x, y))$ is convex wrt w ; because $\max(\cdot)$ is convex
- $\ell(w, (x, y)) \leq \tilde{\ell}(w, (x, y))$ for all $w \in \mathcal{H}$

Then we can compute

$$\tilde{w}_* = \arg \min_{\forall w} \left(\tilde{\ell}_g(w) \right) = \arg \min_{\forall w} \left(\mathbb{E}_{(x,y) \sim g} (\max(0, 1 - y\langle w, x \rangle)) \right)$$

instead of

$$w_* = \arg \min_{\forall x} (R_g(w)) = \arg \min_{\forall x} (\mathbb{E}_{(x,y) \sim g} (1_{(y(w,x) \leq 0)}))$$

Of course by using the surrogate loss instead of the actual one, we introduce some approximation error in the produced output $\tilde{w}_* \neq w_*$.

Remark 33. (Intuitions...) Using a convex surrogate loss function instead the convex one, facilitates computations but introduces extra error to the solution. If $R_g(\cdot)$ is the risk under the non-convex loss, $\tilde{R}_g(\cdot)$ is the risk under the convex surrogate loss, and \tilde{w}_{alg} is the output of the learning algorithm under $\tilde{R}_g(\cdot)$ then we have the upper bound

$$R_g(\tilde{w}_{\text{alg}}) \leq \underbrace{\min_{w \in \mathcal{H}} (R_g(w))}_{\text{I}} + \underbrace{\left(\min_{w \in \mathcal{H}} (\tilde{R}_g(w)) - \min_{w \in \mathcal{H}} (R_g(w)) \right)}_{\text{II}} + \underbrace{\epsilon}_{\text{III}}$$

where term I is the approximation error measuring how well the hypothesis class performs on the generating model, term II is the optimization error due to the use of surrogate loss instead of the actual non-convex one, and term III is the estimation error due to the use of a training set and not the whole generation model.

Handout 0: Learning problem: Definitions, notation, and formulation –A recap

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To get some definitions and set-up about the learning procedure; essentially to formalize what introduced in term 1.

Reading list & references:

- Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.
- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

1. INTRODUCTIONS AND LOOSE DEFINITIONS

Pattern recognition is the automated discovery of patterns and regularities in data $z \in \mathcal{Z}$. **Machine learning (ML)** are statistical procedures for building and understanding probabilistic methods that 'learn'. **ML algorithms** build a (probabilistic/deterministic) model able to make predictions or decisions with minimum human interference and can be used for pattern recognition. **Learning** (or training, estimation) is called the procedure where the ML model is tuned. **Training data** (or observations, sample data set, exemplars) is a set of observables $\{z_i \in \mathcal{Z}\}$ used to tune the parameters of the ML model. **Test set** is a set of available examples/observables $\{z'_i\}$ (different than the training data) used to verify the performance of the ML model for a given a measure of success. **Measure of success** (or performance) is a quantity that indicates how bad the corresponding ML model or Algorithm performs (eg quantifies the failure/error), and can also be used for comparisons among different ML models; eg, **Risk function** or **Empirical Risk Function**. Two main problems in ML are the supervised learning (we focus here) and the unsupervised learning.

Supervised learning problems involve applications where the training data $z \in \mathcal{Z}$ comprises examples of the input vectors $x \in \mathcal{X}$ along with their corresponding target vectors $y \in \mathcal{Y}$; i.e. $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. **Classification problems** are those which aim to assign each input vector x to one of a finite number of discrete categories of y . **Regression problems** are those where the output y consists of one or more continuous variables. All in all, the learner wishes to recover an unknown pattern (i.e. functional relationship) between components $x \in \mathcal{X}$ that serves as inputs and components $y \in \mathcal{Y}$ that act as outputs; i.e. $x \mapsto y$. Hence, \mathcal{X} is the input domain, and \mathcal{Y} is the output domain. The goal of learning is to discover a function which predicts $y \in \mathcal{Y}$ from $x \in \mathcal{X}$.

Unsupervised learning problems involve applications where the training data $z \in \mathcal{Z}$ consists of a set of input vectors $x \in \mathcal{X}$ without any corresponding target values ; i.e. $\mathcal{Z} = \mathcal{X}$. In clustering the goal is to discover groups of similar examples within the data of it is to discover groups of similar examples within the data.

2. (LOOSE) NOTATION IN LEARNING

Definition 1. The learner's output is a function, $h : \mathcal{X} \rightarrow \mathcal{Y}$ which predicts $y \in \mathcal{Y}$ from $x \in \mathcal{X}$. It is also called hypothesis, prediction rule, predictor, or classifier.

Notation 2. We often denote the set of hypothesis as \mathcal{H} ; i.e. $h \in \mathcal{H}$.

Definition 3. Training data set \mathcal{S} of size m is any finite sequence of pairs $((x_i, y_i); i = 1, \dots, m)$ in $\mathcal{X} \times \mathcal{Y}$. This is the information that the learner has assess.

Definition 4. Data generation model $g(\cdot)$ is the probability distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, unknown to the learner that has generated the data.

Definition 5. We denote as $\mathfrak{A}(\mathcal{S})$ the hypothesis (outcome) that a learning algorithm \mathfrak{A} returns given training sample S .

Definition 6. (Loss function) Given any set of hypothesis \mathcal{H} and some domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, a loss function $\ell(\cdot)$ is any function $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}_+$. The purpose of loss function $\ell(h, z)$ is to quantify the “error” for a given hypothesis h and example z –the greater the error the greater its value of the loss.

Example 7. In binary classification problems where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y} = \{0, 1\}$ is discrete, a loss function can be

$$\ell_{0-1}(h, (x, y)) = 1(h(x) = y),$$

Example 8. In regression problems $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and \mathcal{Y} is uncountable, a loss function can be

$$\ell_{\text{sq}}(h, (x, y)) = (h(x) - y)^2$$

Definition 9. (Risk function) The risk function $R_g(h)$ of h is the expected loss of the hypothesis $h \in \mathcal{H}$, w.r.t. probability distribution g over domain Z ; i.e.

$$(2.1) \quad R_g(h) = \mathbb{E}_{z \sim g}(\ell(h, z))$$

Remark 10. In learning, an ideal way to obtain an optimal predictor h^* is to compute the risk minimizer

$$h^* = \arg \min_{\forall} (R_g(h))$$

Example 11. (Cont. Ex. 8) The risk function is $R_g(h) = \mathbb{E}_{z \sim g}(h(x) - y)^2$, and it measures the quality of the hypothesis function $h : \mathcal{X} \rightarrow \mathcal{Y}$ as the expected square difference between the predicted values h and the true target values y at every x .

Remark 12. Computing the risk minimizer may be practically challenging due to the integration w.r.t. the unknown data generation model g involved in the expectation (2.1). Suboptimally, one may resort to the Empirical risk function.

Definition 13. (Empirical risk function) The empirical risk function $\hat{R}_S(h)$ of h is the expectation of loss of h over a given sample $S = (z_1, \dots, z_m) \in \mathcal{Z}^m$; i.e.

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, z_i).$$

Example 14. (Cont. Example 11) Given given sample $S = \{(x_i, y_i); i = 1, \dots, m\}$ the empirical risk function is $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2$.

Example 15. Consider a learning problem where the true data generation distribution (unknown to the learner) is $g(z)$, the statistical model (known to the learner) is given by a sampling distribution $f(y|\theta)$ where the parameter θ is unknown. The goal is to learn θ . If we assume loss function

$$\ell(\theta, z) = \log\left(\frac{g(z)}{f_\theta(z)}\right)$$

then the risk is

$$(2.2) \quad R_g(\theta) = E_{z \sim g}\left(\log\left(\frac{g(z)}{f_\theta(z)}\right)\right) = E_{z \sim g}(\log(g(z))) - E_{z \sim g}(\log(f_\theta(z)))$$

whose minimizer is

$$\theta^* = \arg \min_{\forall \theta} (R_g(\theta)) = \arg \min_{\forall \theta} (E_{z \sim g}(-\log(f_\theta(z))))$$

as the first term in (2.2) is constant. Note that in the Maximum Likelihood Estimation technique the MLE θ_{MLE} is the minimizer of

$$\theta_{MLE} = \arg \min_{\forall \theta} \left(\frac{1}{m} \sum_{i=1}^m (-\log(f_\theta(z_i))) \right)$$

where $S = \{y_1, \dots, y_m\}$ is an IID sample from g . Hence, MLE θ_{MLE} can be considered as the minimizer of the empirical risk $R_S(\theta) = \frac{1}{m} \sum_{i=1}^m (-\log(f_\theta(z_i)))$.

Example 16. (Linear Regression) Consider the multiple linear regression problem $x \mapsto y$ where $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y \in \mathcal{Y} \subseteq \mathbb{R}^d$.

- The hypothesis is a linear function $h : \mathcal{X} \rightarrow \mathcal{Y}$ (that learner wishes to learn) to approximate mapping $x \mapsto y$. The hypothesis set $\mathcal{H} = \{\langle w, x \rangle \mapsto y : w \in \mathbb{R}^d\}$. We can use the loss $\ell(h, (x, y)) = (h(x) - y)^2$.
- Equivalently, learning problem can be set differently because the predictor (linear function) is parametrized by $w \in \mathbb{R}^d$ as $\langle w, x \rangle \mapsto y$. Set $\mathcal{H} = \{w \in \mathbb{R}^d\}$. The set of examples is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The loss is $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$.

Handout 2: Gradient descent

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce gradient descent, its motivation, description, practical tricks, analysis in the convex scenario, and implementation.

Reading list & references:

- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
– Ch. 14.1 Gradient Descent

Further reading

- Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.

1. MOTIVATIONS

Note 1. Consider a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$. Learning may involve the computation of the minimizer $h^* \in \mathcal{H}$, where \mathcal{H} is a class of hypotheses, of the empirical risk function (ERF) $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h, z_i)$ given a finite sample $\{z_i; i = 1, \dots, n\}$ generated from the data generating model $g(\cdot)$ and using loss $\ell(\cdot)$; that is

$$(1.1) \quad h^* = \arg \min_{\forall h \in \mathcal{H}} (\hat{R}(h)) = \arg \min_{\forall h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \ell(h, z_i) \right)$$

If analytical minimization of (1.1) is impossible or impractical, numerical procedures can be applied; eg Gradient Descent (GD) algorithms. Such approaches introduce numerical errors in the solution.

2. DESCRIPTION

Notation 2. For the sake of notation simplicity and generalization, we will present Gradient Descent (GD) in the following minimization problem

$$(2.1) \quad w^* = \arg \min_{\forall w \in \mathcal{H}} (f(w))$$

where here $f : \mathbb{R}^d \rightarrow \mathbb{R}$, and $w \in \mathcal{H} \subseteq \mathbb{R}^d$; $f(\cdot)$ is the function to be minimized, e.g., $f(\cdot)$ can be an empirical risk function $\hat{R}(\cdot)$.

Assumption 3. Assume (for now) that $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a differentiable function.

Definition 4. Given a per-specified learning rate $\eta_t > 0$, the Gradient Descent (GD) algorithm for the solution of the minimization problem (2.1) is given in Algorithm 1

Algorithm 1 Gradient descent algorithm with learning rate η_t

For $t = 1, 2, 3, \dots$ iterate:

(1) compute

$$(2.2) \quad w^{(t+1)} = w^{(t)} - \eta_t \nabla f(w^{(t)})$$

(2) terminate if a termination criterion is satisfied, e.g.

If $t \geq T_{\max}$ then STOP

where

$$\nabla f(w) = \left(\frac{\partial}{\partial x_1} f(x), \dots, \frac{\partial}{\partial x_d} f(x) \right)^T \Big|_{x=w}$$

is the gradient of f at w .

Remark 5. Interpreting GD, GD produces a chain $\{w^{(t)}\}$ that drifts towards the minimum w^* . It evolves directed towards the opposite direction than that of the gradient $\nabla f(\cdot)$ and at a rate controlled by the learning rate η_t .

Remark 6. For more intuitive explanation, consider the (1st order) Taylor polynomial for the approximation of $f(w)$ in a small area around u (i.e. $\|v - u\| = \text{small}$)

$$f(u) \approx P(u) = f(w) + \langle u - w, \nabla f(w) \rangle$$

Assuming convexity for f , it is

$$(2.3) \quad f(u) \geq \underbrace{f(w) + \langle u - w, \nabla f(w) \rangle}_{=P(u;w)}$$

See
Handout 1

meaning that P lower bounds f . Hence we could design an updating mechanism producing $w^{(t+1)}$ which is nearby $w^{(t)}$ (small steps) and which minimize the linear approximation $P(w)$ of $f(w)$ at $w^{(t)}$

$$(2.4) \quad P(w; w^{(t)}) = f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle.$$

while hoping that this mechanism would push the produced chain $\{w^{(t)}\}$ towards the minimum because of (2.3). Hence we could recursively minimize the linear approximation (2.4) and the distance between the current state $w^{(t)}$ and the next w value to produce $w^{(t+1)}$; namely

$$(2.5) \quad \begin{aligned} w^{(t+1)} &= \arg \min_{\forall w} \left(\frac{1}{2} \|w - w^{(t)}\|^2 + \eta P(w; w^{(t)}) \right) \\ &= \arg \min_{\forall w} \left(\frac{1}{2} \|w - w^{(t)}\|^2 + \eta \left(f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t+1)}) \rangle \right) \right) \\ &= w^{(t)} - \eta \nabla f(w^{(t)}) \end{aligned}$$

where parameter $\eta > 0$ controls the trade off in (2.5).

Remark 7. Given T GD algorithm iterations, the output of GD can be (but not exclusively),

(1) the average (after discarding the first few iterations of $w^{(t)}$ for stability reasons)

$$(2.6) \quad w_{\text{GD}}^{(T)} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$$

(2) or the best value discovered

$$w_{\text{GD}}^{(T)} = \arg \min_{\forall w_t} \left(f(w^{(t)}) \right)$$

(3) or the last value discovered

$$w_{\text{GD}}^{(T)} = w^{(T)}$$

Note 8. GD output converges to a local minimum, $w_{\text{GD}}^{(T)} \rightarrow w_*$ (in some sense), under different sets of regularity conditions (some are weaker other stronger). Section 4 has a brief analysis.

Remark 9. The parameter η_t is called learning rate, step size, or gain. It determines the size of the steps GD takes to reach a (local) minimum. $\{\eta_t\}$ is a non-negative sequence and it is chosen by the practitioner. In principle, regularity conditions (Note 8) often imply restrictions on the decay of $\{\eta_t\}$ which guide the practitioner to parametrize it properly. Some popular choices of learning rate η_t are:

- (1) constant; $\eta_t = \eta$, for where $\eta > 0$ is a small value. The rationale is that GD chain $\{w_t\}$ performs constant small steps towards the (local) minimum w_* and then oscillate around it.
 - (2) decreasing and converging to zero; $\eta_t \downarrow$ with $\lim_{t \rightarrow \infty} \eta_t = 0$. E.g. $\eta_t = (\frac{C}{t})^\varsigma$ where $\varsigma \in [0.5, 1]$ and $C > 0$. The rationale is that GD algorithm starts by performing larger steps (controlled by C) at the begining to explore the area for discovering possible minima. Also it reduces the size of those steps with the iterations (controled by ς) such that eventually when the chain $\{w_t\}$ is close to a possible minimum w_* value to converge and do not overshoot.
 - (3) decreasing and converging to a tiny value τ_* ; $\eta_t \downarrow$ with $\lim_{t \rightarrow \infty} \eta_t = \tau_*$. E.g. $\eta_t = (\frac{C}{t})^\varsigma + \tau_*$ with $\varsigma \in (0.5, 1]$, $C > 0$, and $\tau_* \approx 0$. Same as previously, but the algorithm aims at oscillating around the detected local minimum.
 - (4) constant until an iteration T_0 and then decreasing; Eg $\eta_t = \left(\frac{C}{\max(t, T_0)} \right)^\varsigma$ with $\varsigma \in [0.5, 1]$ and $C > 0$, and $T_0 < T$. The rationale is that at the first stage of the iterations (when $t \leq T_0$) the algorithm may need a constant large steps for a significant number of iterations T_0 in order to explore the domain; and hence in order for the chain $\{w_t\}$ to reach the area around the (local) minimum w_* . In the second stage, hoping that the chain $\{w_t\}$ may be in close proximity to the (local) minimum w_* the algorithm progressively performs smaller steps to converge towards the minimum w_* . The first stage ($t \leq T_0$) is called burn-in; the values $\{w_t\}$ produced during the burn-in ($t \leq T_0$) are often discarded/ignored from the output of the GD algorithm.
- Parameters $C, \varsigma, \tau_*, T_0$ may be chosen based on pilot runs.

Remark 10. There are several practical termination criteria that can be used in GD Algorithm 1(step 2). They aim to terminate the recursion in practice. Some popular termination criteria are

- (1) terminate when the gradient is sufficiently close to zero; i.e. if $\|\nabla f(w^{(t)})\| \leq \epsilon$ for some pre-specified tiny $\epsilon > 0$ then STOP
- (2) terminate when the chain $w^{(t)}$ does not change; i.e. if $\|w^{(t+1)} - w^{(t)}\| \leq \epsilon \|w^{(t)}\|$ for some pre-specified tiny $\epsilon > 0$ then STOP
- (3) terminate when a pre-specified number of iterations T is performed; i.e. if $t \geq T$ then STOP

Here (1) may be deceive if the chain is in a flat area, (2) may be deceived if the learning rate become too small, (3) is obviously a last resort.

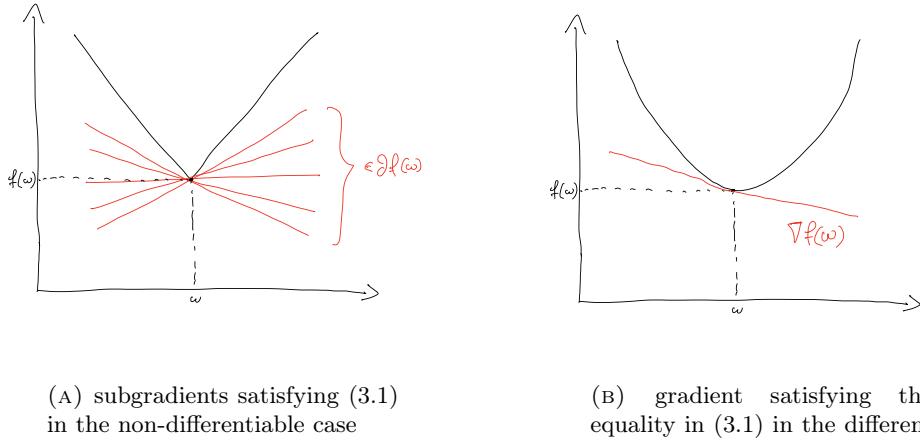
3. GD FOR NON-DIFFERENTIABLE FUNCTIONS (USING SUB-GRADIENTS)

Note 11. In several learning problems the function to be minimized is not differentiable. GD can be extended to address such problems with the use of subgradients.

Definition 12. Vector v is called subgradient of a function $f : S \rightarrow \mathbb{R}$ at $w \in S$ if

$$(3.1) \quad \forall u \in S, \quad f(u) \geq f(w) + \langle u - w, v \rangle$$

Note 13. Essentially there may be more than one subgradients of the function at a specific point. As seen by (3.1), subgradients are the slopes of all the lines passing through the point $(w, f(w))$ and been under the function $f(\cdot)$.



Notation 14. The set of subgradients of function $f : S \rightarrow \mathbb{R}$ at $w \in S$ is denoted by $\partial f(w)$.

Definition 15. The Gradient Descent algorithm using subgradients in non-differentiable cases, results by replacing the gradient $\nabla f(w^{(t)})$ in (2.2) with any of subgradient v_t from the set of subgradients $\partial f(w^{(t)})$ at $w^{(t)}$; namely

$$(3.2) \quad w^{(t+1)} = w^{(t)} - \eta_t v_t; \quad \text{where } v_t \in \partial f(w^{(t)})$$

3.1. Construction of subgradient.

Note 16. We discuss how to construct subgradients in practice.

Fact 17. *Some properties of subgradient sets that help for their construction*

- (1) *If function $f : S \rightarrow \mathbb{R}$ is differentiable at w then the only subgradient of f at w is the gradient $\nabla f(w)$, and (3.1) is equality; i.e. $\partial f(w) = \{\nabla f(w)\}$.*
- (2) *for constants α, β and convex function $f(\cdot)$, it is*

$$\partial(\alpha f(w) + \beta) = \alpha(\partial f(w)) = \{\alpha v : v \in \partial f(w)\}$$

- (3) *for convex functions $f(\cdot)$ and $g(\cdot)$, it is*

$$\partial(f(w) + g(w)) = \partial f(w) + \partial g(w) = \{v + u : v \in \partial f(w), \text{ and } u \in \partial g(w)\}$$

Example 18. Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}_+$ with $f(w) = |w| = \begin{cases} w & w \geq 0 \\ -w & w < 0 \end{cases}$. Find the set of subgradients $\partial f(w)$ for each $w \in \mathbb{R}$.

Solution. Using Fact 17, it is $\partial f(w) = 1$ for $w > 0$ and $\partial f(w) = -1$ for $w < 0$ as f is differentiable for $x \neq 0$. At $x = 0$, f is not differentiable; hence from condition (3.1) it is

$$\forall u \in \mathbb{R}, \quad |u| \geq |0| + (u - 0)v$$

which is satisfied for $v \in [-1, 1]$. Hence,

$$\partial f(w) = \begin{cases} \{-1\} & , w < 0 \\ [-1, 1] & , w = 0 \\ \{1\} & , w > 0 \end{cases}$$

4. ANALYSIS OF GRADIENT DESCENT

Assumption 19. *For the sake of the analysis of the GD, let us consider:*

- (1) *constant learning rate $\eta_t = \eta$,*
- (2) *GD output $w_{GD}^{(T)} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$*

Also as we will see, function $f(\cdot)$ should be convex and Lipschitz in order for the following results to hold

Notation 20. w^* is the minimizer in (2.1).

Note 21. We consider Lemma 22 as given Fact.

Lemma 22. *Let $\{v_t; t = 1, \dots, T\}$ be a sequence of vectors. Any algorithm with $w^{(1)} = 0$ and $w^{(t+1)} = w^{(t)} - \eta v_t$ for $t = 1, \dots, T$ satisfies*

$$(4.1) \quad \sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

Proof. Omitted; see the 2nd reference. \square

Note 23. To find an upper bound of the GD error, we try to bound the error $f(w_{\text{GD}}^{(T)}) - f(w^*)$ with purpose to use Lemma 22.

Proposition 24. Consider the minimization problem (2.1). Given Assumptions 19, the error can be bounded as

$$(4.2) \quad f(w_{\text{GD}}^{(T)}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle$$

where $v_t \in \partial f(w^{(t)})$. If $f(\cdot)$ is differentiable then $v_t = f'(w^{(t)})$

Proof. It is¹

$$\begin{aligned} f(w_{\text{GD}}^{(T)}) - f(w^*) &= f\left(\frac{1}{T} \sum_{t=1}^T w_t\right) - f(w^*) \\ (4.3) \quad &\leq \frac{1}{T} \sum_{t=1}^T (f(w_t) - f(w^*)) && \text{(by Jensen's inequality)} \\ &\leq \frac{1}{T} \sum_{t=1}^T \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle && \text{(by convexity of } f(\cdot)) \end{aligned}$$

\square

Note 25. The following provides conditions under which the gradient (or sub-gradient) is bounded. This is necessary in order to bound (4.2) with (4.1) in a meaningful manner.

Proposition 26. ² $f : S \rightarrow \mathbb{R}$ is ρ -Lipschitz over an open convex set S if and only if for all $w \in S$ and $v \in \partial f(w)$ it is $\|v\| \leq \rho$.

Proof. \implies Let $f : S \rightarrow \mathbb{R}$ be ρ -Lipschitz over convex set S , $w \in S$ and $v \in \partial f(w)$.

- Since S is open we get that there exist $\epsilon > 0$ such as $u := w + \epsilon \frac{v}{\|v\|}$ where $u \in S$. So $\langle u - w, v \rangle = \epsilon \|v\|$ and $\|u - w\| = \epsilon$.
- From the subgradient definition we get

$$f(u) - f(w) \geq \langle u - w, v \rangle = \epsilon \|v\|$$

- From the Lipschitzness of $f(\cdot)$ we get

$$f(u) - f(w) \leq \rho \|u - w\| = \rho \epsilon$$

Therefore $\|v\| \leq \rho$.

For \impliedby see Exercise 4 in the Exercise sheet. \square

¹Jensen's inequality for convex $f(\cdot)$ is $E(f(x)) \leq f(E(x))$

²If this was a Homework there would be a Hint:

- If S is open there exist $\epsilon > 0$ such as $u = w + \epsilon \frac{v}{\|v\|}$ such as $u \in S$

Note 27. The following summarizes Lemma 22, and Propositions 24, 26 with respect to the GD algorithm.

Proposition 28. *Let $f(\cdot)$ be a convex and ρ -Lipschitz function. If we run GD algorithm of f with learning rate $\eta > 0$ for T steps the output $w_{\text{GD}}^{(T)} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$ satisfies*

$$f\left(w_{\text{GD}}^{(T)}\right) - f(w^*) \leq \frac{\|w^*\|^2}{2\eta T} + \frac{\eta}{2} \frac{1}{T} \sum_{t=1}^T \|\nabla f(w^{(t)})\|^2$$

where $\|\nabla f(\cdot)\| \leq \rho$.

Solution. Straightforward from Lemma 22, and Propositions 24, 26.

Note 29. The following shows that a given learning rate depending on the iteration t , we can reduce the upper bound of the error as well as find the number of required iterations to achieve convergence.

Proposition 30. *(Cont Prop. 28) Let $f(\cdot)$ be a convex and ρ -Lipschitz function, and let $\mathcal{H} = \{w \in \mathbb{R} : \|w\| \leq B\}$. Assume we run GD algorithm of $f(\cdot)$ with learning rate $\eta_t = \sqrt{\frac{B^2}{\rho^2 T}}$ for T steps, and output $w_{\text{GD}}^{(T)} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$. Then*

(1) *upper bound on the sub-optimality is*

$$(4.4) \quad f\left(w_{\text{GD}}^{(T)}\right) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}$$

(2) *a given level off accuracy ε such that $f\left(w_{\text{GD}}^{(T)}\right) - f(w^*) \leq \varepsilon$ can be achieved after T iterations*

$$T \geq \frac{B^2 \rho^2}{\varepsilon^2}.$$

Proof. Part 1 is a simple substitution from Proposition 28, and part 2 is implied from part 1. \square

Note 31. The result on Proposition 30 heavily relies on setting suitable values for B and ρ which is rather a difficult task to be done in very complicated learning problems (e.g., learning a neural network).

Remark 32. The above results from the analysis of the GD also hold for the GD with subgradients; just replace $\nabla f(\cdot)$ with any v_t such that $v_t \in \partial f(\cdot)$.

5. EXAMPLES³

Example 33. Consider the simple Normal linear regression problem where the dataset $\{z_i = (y_i, x_i)\}_{i=1}^n \in \mathcal{D}$ is generated from a Normal data generating model

$$(5.1) \quad \begin{pmatrix} y_i \\ x_i \end{pmatrix} \stackrel{\text{iid}}{\sim} N\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{bmatrix} \sigma_y^2 & \rho\sqrt{\sigma_y^2\sigma_x^2} \\ \rho\sqrt{\sigma_y^2\sigma_x^2} & \sigma_x^2 \end{bmatrix}\right)$$

³Code is available in https://github.com/georgios-stats/Machine_Learning_and_Neural_Networks_III_Epiphanys_2023/tree/main/Lecture_handouts/code/02.Gradient_descent/example_1.R

for $i = 1, \dots, n$. Consider a hypothesis space \mathcal{H} of linear functions $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ with $h(w) = w_1 + w_2 x$. The exact solution (which we pretend we do not know) is given as

$$(5.2) \quad \begin{pmatrix} w_1^* \\ w_2^* \end{pmatrix} = \begin{pmatrix} \mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x \\ \rho \frac{\sigma_y}{\sigma_x} \end{pmatrix}.$$

To learn the optimal $w^* = (w_1^*, w_2^*)^\top$, we consider a loss $\ell(w, z_i = (x_i, y_i)^\top) = (y_i - [w_1 + w_2 x_i])^2$, which leads to the minimization problem

$$w^* = \arg \min_{\forall w} \left(\hat{R}_{\mathcal{D}}(w) \right) = \arg \min_{\forall w} \left(\frac{1}{n} \sum_{i=1}^n (y_i - w_1 - w_2 x_i)^2 \right)$$

The GD Algorithm 1 with learning rate η is

For $t = 1, 2, 3, \dots$ iterate:

(1) compute

$$(5.3) \quad w^{(t+1)} = w^{(t)} - \eta v_t,$$

$$\text{where } v_t = \begin{pmatrix} 2w_1^{(t)} + 2w_2^{(t)} \bar{x} - 2\bar{y} \\ 2w_1^{(t)} \bar{x} + 2w_2^{(t)} \bar{x}^2 - 2y^\top x \end{pmatrix}$$

(2) terminate if a termination criterion is satisfied, e.g.

If $t \geq T$ then STOP

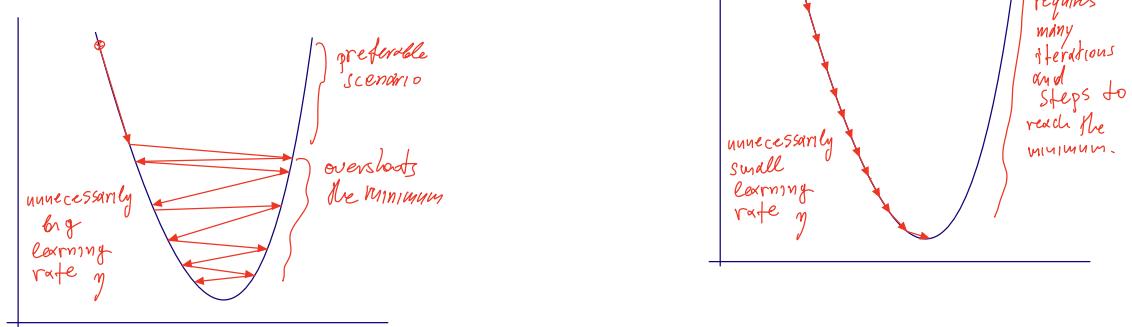
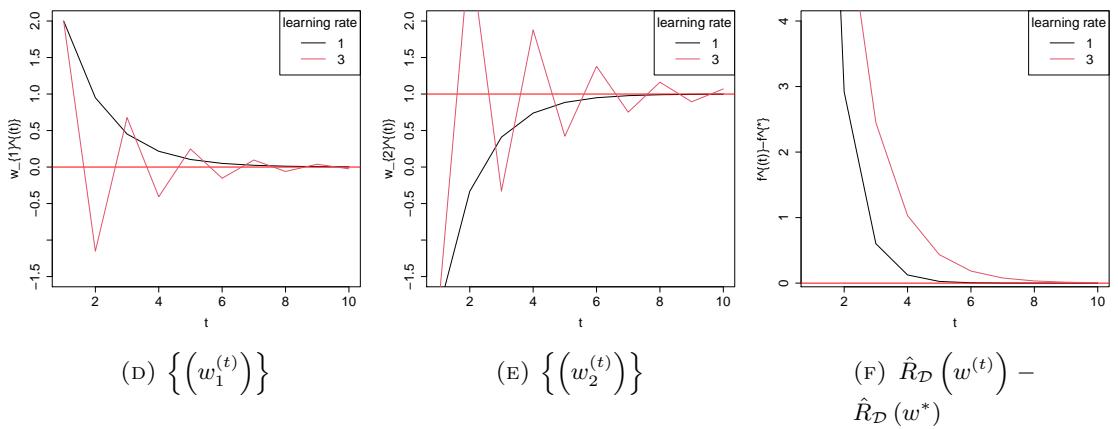
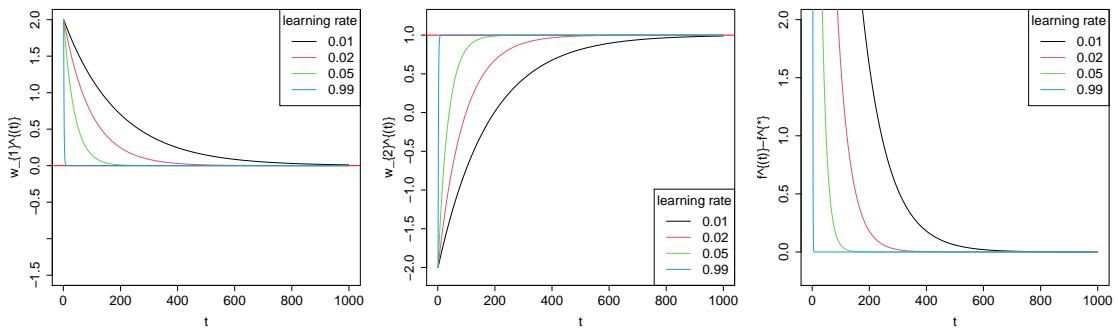
This is because $\hat{R}_{\mathcal{D}}(w)$ is differentiable in \mathbb{R}^2 so $\partial \hat{R}_{\mathcal{D}}(w) = \{\nabla \hat{R}_{\mathcal{D}}(w)\}$ and because

$$\nabla \hat{R}_{\mathcal{D}}(w) = \begin{pmatrix} \frac{d}{dw_1} \hat{R}_{\mathcal{D}}(w) \\ \frac{d}{dw_2} \hat{R}_{\mathcal{D}}(w) \end{pmatrix} = \dots = \begin{pmatrix} 2w_1^{(t)} + 2w_2^{(t)} \bar{x} - 2\bar{y} \\ 2w_1^{(t)} \bar{x} + 2w_2^{(t)} \bar{x}^2 - 2y^\top x \end{pmatrix}$$

Consider data size $n = 100$, and parameters $\rho = 0.2$, $\sigma_y^2 = 1$ and $\sigma_x^2 = 1$. Then the real value (5.2) that I need to learn equals to $w^* = (0, 1)^\top$. Consider a GD seed $w_0 = (2, -2)$, and total number of iterations $T = 1000$.

Figures 5.1a, 5.1b, and 5.1c present trace plots of the chain $\{(w^{(t)})\}$ and error $\hat{R}_{\mathcal{D}}(w^{(t)}) - \hat{R}_{\mathcal{D}}(w^*)$ produced by running GD for $T = 1000$ total iterations and for different (each time) constant learning rates $\eta \in \{0.01, 0.02, 0.05, 0.99\}$. We observe that the larger learning rates under consideration were able to converge faster to the minimum w^* . This is because they perform larger steps and can learn faster -this is not a panacea.

Figures 5.1d, 5.1e, and 5.1f present trace plots of the chain $\{(w^{(t)})\}$, and of the error $\hat{R}_{\mathcal{D}}(w^{(t)}) - \hat{R}_{\mathcal{D}}(w^*)$ produced by running GD for $T = 1000$ total iterations and for learning rate $\eta = 1.0$ (previously considered) and a very big learning rate $\eta = 3.0$. We observe that the very big learning rate $\eta = 3.0$ presents slower convergence to the minimum w^* . This is because it creates unreasonably big steps in (2.2) that the produced chain overshoots the global minimum. See the cartoon in Figures 5.1a and 5.1b.



(A) Unnecessarily large learning rate

(B) Unnecessarily small learning rate

Handout 3: Stochastic gradient descent

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce the stochastic gradient descent (motivation, description, practical tricks, analysis in the convex scenario, and implementation).

Reading list & references:

- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
 - Ch. 14.3 Stochastic Gradient Descent (SGD), 14.5 Variants, 14.5 Learning with SGD

Further reading

- Bottou, L. (2012). Stochastic gradient descent tricks. In Neural networks: Tricks of the trade (pp. 421-436). Springer, Berlin, Heidelberg.

1. MOTIVATIONS FOR STOCHASTIC GRADIENT DESCENT

Problem 1. Consider a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$. Learning may involve the computation of the minimizer $w^* \in \mathcal{H}$, where \mathcal{H} is a class of hypotheses, of the risk function (RF) $R(w) = \mathbb{E}_{z \sim g}(\ell(w, z))$ given an unknown data generating model $g(\cdot)$ and using a known tractable loss $\ell(\cdot, \cdot)$; that is

$$(1.1) \quad w^* = \arg \min_{w \in \mathcal{H}} (R_g(w)) = \arg \min_{w \in \mathcal{H}} (\mathbb{E}_{z \sim g}(\ell(w, z)))$$

Remark 2. Gradient descent (GD) cannot be directly utilized to address Problem 1 (i.e., minimize the Risk function) because g is unknown, and because (1.1) involves an integral which may be computationally intractable. Instead it aims to minimize the ERF $\hat{R}(w) = \frac{1}{n} \sum_{i=1}^n \ell(w, z_i)$ which ideally is used as a proxy when data size n is big (big-data).

Remark 3. The implementation of GD may be computationally impractical even in problems where we need to minimize an ERF $\hat{R}_n(w)$ if we have big data ($n \approx$ big). This is because GD requires the recursive computation of the exact gradient $\nabla \hat{R}_n(w) = \frac{1}{n} \sum_{i=1}^n \nabla \ell(w, z_i)$ using all the data $\{z_i\}$ at each iteration. That may be too slow.

Remark 4. Stochastic gradient descent (SGD) aims at solving (1.1), and overcoming the issues in Remarks 2 & 3 by using an unbiased estimator of the actual gradient (or some sub-gradient) based on a sample properly drawn from g .

2. STOCHASTIC GRADIENT DESCENT

2.1. Description.

Notation 5. For the sake of notation simplicity and generalization, we present Stochastic Gradient Descent (SGD) in the following minimization problem

$$(2.1) \quad w^* = \arg \min_{\forall w \in \mathcal{H}} (f(w))$$

where here $f : \mathbb{R}^d \rightarrow \mathbb{R}$, and $w \in \mathcal{H} \subseteq \mathbb{R}^d$; $f(\cdot)$ is the unknown function to be minimized, e.g., $f(\cdot)$ can be the risk function $R_g(w) = \mathbb{E}_{z \sim g}(\ell(w, z))$.

Algorithm 6. *Stochastic Gradient Descent (SGD) with learning rate $\eta_t > 0$ for the solution of the minimization problem (2.1)*

For $t = 1, 2, 3, \dots$ iterate:

- (1) compute

$$(2.2) \quad w^{(t+1)} = w^{(t)} - \eta_t v_t,$$

where v_t is a random vector such that $E(v_t | w^{(t)}) \in \partial f(w^{(t)})$

- (2) terminate if a termination criterion is satisfied, e.g.

If $t \geq T_{\max}$ then STOP

Remark 7. If f is differentiable at $w^{(t)}$, it is $\partial f(w^{(t)}) = \{\nabla f(w^{(t)})\}$. Hence v_t is such as $E(v_t | w^{(t)}) = \nabla f(w^{(t)})$ in Algorithm 6 step 1.

Note 8. Assume f is differentiable (for simplicity). To compare SGD with GD, we can re-write (2.2) in the SGD Algorithm 6 as

$$(2.3) \quad w^{(t+1)} = w^{(t)} - \eta_t [\nabla f(w^{(t)}) + \xi_t],$$

where

$$\xi_t := v_t - \nabla f(w^{(t)})$$

represents the (observed) noise introduced in (2.2) by using a random realization of the exact gradient.

Remark 9. Given T SGD algorithm iterations, the output of SGD can be (but not exclusively)

- (1) the average (after discarding the first few iterations of $w^{(t)}$ for stability reasons)

$$(2.4) \quad w_{\text{SGD}}^{(T)} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$$

- (2) or the best value discovered

$$w_{\text{SGD}}^{(T)} = \arg \min_{\forall w_t} (f(w^{(t)}))$$

- (3) or the last value discovered

$$w_{\text{SGD}}^{(T)} = w^{(T)}$$

Note 10. SGD output converges to a local minimum, $w_{\text{SGD}}^{(T)} \rightarrow w_*$ (in some sense), under different sets of regularity conditions. Section 4 has a brief analysis. To achieve this, Conditions 11 on the learning rate are rather inevitable and should be satisfied.

Condition 11. Regarding the learning rate (or gain) $\{\eta_t\}$ should satisfy conditions

- (1) $\eta_t \geq 0$,
- (2) $\sum_{t=1}^{\infty} \eta_t = \infty$
- (3) $\sum_{t=1}^{\infty} \eta_t^2 < \infty$

Remark 12. The popular learning rates $\{\eta_t\}$ in Remark 9 in Handout 2 satisfy Condition 11 and hence can be used in SGD too. Once parametrized, η_t can be tuned based on pilot runs using a reasonably small number of data.

Remark 13. Intuition on Condition 11. Assume that v_t is bounded. Condition 11(3) aims at reducing the effect of the randomness in v_t (introduced noise ξ_t) because it implies $\eta_t \searrow 0$ as $t \rightarrow \infty$; if this was not the case then

$$w^{(t+1)} - w^{(t)} = -\eta_t v_t \rightarrow 0$$

may not be satisfied and the chain $\{w^{(t)}\}$ may not converge. Condition 11(2) prevents η_t from reducing too fast and allows the generated chain $\{w^{(t)}\}$ to be able to converge. E.g., after t iterations

$$\begin{aligned} \|w^{(t)} - w^*\| &= \|w^{(t)} \pm w^{(0)} - w^*\| \geq \|w^{(0)} - w^*\| - \|w^{(t)} - w^{(0)}\| \\ &\geq \|w^{(0)} - w^*\| - \sum_{t=0}^{\infty} \|w^{(t+1)} - w^{(t)}\| = \|w^{(0)} - w^*\| - \sum_{t=0}^{T-1} \|\eta_t v_t\| \end{aligned}$$

However if it was $\sum_{t=1}^{\infty} \eta_t < \infty$ it would be $\sum_{t=0}^{\infty} \|\eta_t v_t\| < \infty$ and hence $w^{(t)}$ would never converge to w^* if the seed $w^{(0)}$ is far enough from w^* .

3. STOCHASTIC GRADIENT DESCENT WITH PROJECTION

Remark 14. Consider the scenario in Problem 1 where the learning problem requires to discover w^* in the restricted/bounded set \mathcal{H} . Assume the function to be minimized is convex in the restricted hypothesis set \mathcal{H} , e.g. $\mathcal{H} = \{w : \|w\| \leq B\}$, but non-convex in \mathbb{R}^d . Direct implementation of vanilla SGD (Algorithm 6) may produce a chain stepping out \mathcal{H} and hence an output $w_{\text{SGD}} \notin \mathcal{H}$. SGD can be modified to address this issue, as in Algorithm 15, by including a projection step guarantying $w \in \mathcal{H}$.

Algorithm 15. *Stochastic Gradient Descent with learning rate $\eta_t > 0$ and with projection in \mathcal{H} for problem in (2.1)*

For $t = 1, 2, 3, \dots$ iterate:

(1) compute

$$(3.1) \quad w^{(t+\frac{1}{2})} = w^{(t)} - \eta_t v_t,$$

where v_t is a random vector such that $E(v_t|w^{(t)}) \in \partial f(w^{(t)})$

(2) compute

$$(3.2) \quad w^{(t+1)} = \arg \min_{w \in \mathcal{H}} \left(\|w - w^{(t+\frac{1}{2})}\| \right)$$

(3) terminate if a termination criterion is satisfied

4. ANALYSIS OF SGD (ALGORITHM 6)

Note 16. Recall that the stochasticity of SGD comes from the stochastic sub-gradients $\{v_t\}_{\forall t}$; hence the expectations below are under these random vectors' distributions.

Theorem 17. *Let $f(\cdot)$ be a convex function. If we run SGD algorithm of f with learning rate $\eta_t > 0$ for T steps, the output $w_{\text{GD}}^{(T)} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$ satisfies*

$$(4.1) \quad E \left(f \left(w_{\text{SGD}}^{(T)} \right) \right) - f(w^*) \leq \frac{\|w^*\|^2}{2\eta T} + \frac{\eta}{2} \sum_{t=1}^T E \|v_t\|^2$$

Proof. Let $v_{1:t} = (v_1, \dots, v_t)$. By Jensens' inequality (or see (4.3) in Handout 2)

$$(4.2) \quad E \left(f \left(w_{\text{SGD}}^{(T)} \right) - f(w^*) \right) \leq E \left(\frac{1}{T} \sum_{t=1}^T \left(f \left(w^{(t)} \right) - f(w^*) \right) \right) = \frac{1}{T} \sum_{t=1}^T E \left(f \left(w^{(t)} \right) - f(w^*) \right)$$

I will try to use Lemma 22 from Handout 2, hence I need to show

$$(4.3) \quad E \left(f \left(w^{(t)} \right) - f(w^*) \right) \leq E \left(\langle w^{(t)} - w^*, v_t \rangle \right)$$

where the expectation is under $v_{1:T}$. It is

$$\begin{aligned} E_{v_{1:T}} \left(\langle w^{(t)} - w^*, v_t \rangle \right) &= E_{v_{1:t}} \left(\langle w^{(t)} - w^*, v_t \rangle \right) \\ &= E_{v_{1:t-1}} \left(E_{v_{1:t}} \left(\langle w^{(t)} - w^*, v_t \rangle | v_{1:t-1} \right) \right) \quad (\text{law of total expectation}) \end{aligned}$$

But $w^{(t)}$ is fully determined by $v_{1:t-1}$, (see (2.2)) so

$$E_{v_{1:t-1}} \left(E_{v_{1:t}} \left(\langle w^{(t)} - w^*, v_t \rangle | v_{1:t-1} \right) \right) = E_{v_{1:t-1}} \left(\langle w^{(t)} - w^*, E_{v_{1:t}} (v_t | v_{1:t-1}) \rangle \right)$$

As $w^{(t)}$ is fully determined by $v_{1:t-1}$ then $\mathbb{E}_{v_{1:t}}(v_t|v_{1:t-1}) = \mathbb{E}_{v_{1:t}}(v_t|w^{(t)}) \in \partial f(w^{(t)})$, hence $\mathbb{E}_{v_{1:t}}(v_t|v_{1:t-1})$ is a sub-gradient. By sub-gradient definition

$$\begin{aligned} \mathbb{E}_{v_{1:t-1}}(\langle w^{(t)} - w^*, \mathbb{E}_{v_{1:t}}(v_t|v_{1:t-1}) \rangle) &\geq \mathbb{E}_{v_{1:t-1}}(f(w^{(t)}) - f(w^*)) \\ (4.4) \quad &= \mathbb{E}_{v_{1:T}}(f(w^{(t)}) - f(w^*)) \end{aligned}$$

Hence combining (4.4), (4.3), and (4.2)

$$\mathbb{E}(f(w_{\text{SGD}}^{(T)}) - f(w^*)) \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}(\langle w^{(t)} - w^*, v_t \rangle)$$

Then Lemma 22 in Handout 2 implies

$$\mathbb{E}(f(w_{\text{SGD}}^{(T)}) - f(w^*)) \leq \mathbb{E}\left(\frac{1}{T} \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \frac{1}{T} \sum_{t=1}^T \|v_t\|^2\right) = \frac{\mathbb{E}\|w^*\|^2}{2\eta T} + \frac{\eta}{2} \frac{1}{T} \sum_{t=1}^T \mathbb{E}\|v_t\|^2$$

□

Remark 18. The upper bound in (4.1) depends on the variation of v_t as

$$(4.5) \quad \mathbb{E}\|v_t\|^2 = \sum_{j=1}^d \text{Var}(v_{t,j}) + \sum_{j=1}^d (\mathbb{E}(v_{t,j}))^2$$

where d is the dimension of $v_t = (v_{t,1}, \dots, v_{t,d})^\top$. The second term on the right hand side of (4.5) is constant as $v_{t,j}$ is the unbiased estimator of the sub-gradient by construction.

Proposition 19. (*Cont. Theorem 17*) Let $f(\cdot)$ be a convex function, and let $\mathcal{H} = \{w \in \mathbb{R} : \|w\| \leq B\}$. Let $\mathbb{E}\|v_t\|^2 \leq \rho^2$. Assume we run SGD algorithm of $f(\cdot)$ with learning rate $\eta_t = \sqrt{\frac{B^2}{\rho^2 T}}$ for T steps, and output $w_{\text{SGD}}^{(T)} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$. Then

(1) upper bound on the sub-optimality is

$$(4.6) \quad \mathbb{E}(f(w_{\text{SGD}}^{(T)}) - f(w^*)) \leq \frac{B\rho}{\sqrt{T}}$$

(2) a given level off accuracy ε such that $\mathbb{E}(f(w_{\text{SGD}}^{(T)}) - f(w^*)) \leq \varepsilon$ can be achieved after T iterations

$$T \geq \frac{B^2 \rho^2}{\varepsilon^2}.$$

Proof. It follows from 17. □

Remark 20. Lemma 22 in Handout 2 holds even when projection steps are added in the vanilla SGD algorithm, hence the above analysis holds for the SGD with projections (Algorithm 15) too.

5. IMPLEMENTATION OF SGD IN THE LEARNING PROBLEM 1

Note 21. Below we show SGD can be implemented in the learning Problem 1. Proposition 22 allows a convenient implementation of SGD to the learning problem (Problem 1).

Proposition 22. For a randomly drawn $z \sim g(\cdot)$, the sub-gradient v of $\ell(w, z)$ at point w is an unbiased estimator of the sub-gradient of the risk $R_g(w)$ at point w .

Proof. Let v be a sub-gradient of $\ell(w, z)$ at point w , then

$$(5.1) \quad \ell(u, z) - \ell(w, z) \geq \langle u - w, v \rangle$$

It is

$$\begin{aligned} R_g(u) - R_g(w) &= \mathbb{E}_{z \sim g} (\ell(u, z) - \ell(w, z) | w) \geq \mathbb{E}_{z \sim g} (\langle u - w, v \rangle | w) \\ &= \langle u - w, \mathbb{E}_{z \sim g} (v | w) \rangle \end{aligned}$$

Hence, by definition, v is such that $\mathbb{E}_{z \sim g} (v | w)$ is a sub-gradient of $R_g(w)$. \square

Example 23. Consider that loss ℓ is differentiable wrt w . If $v = \nabla_w \ell(w, z)$, then

$$\mathbb{E}_{z \sim g} (v | w) = \nabla_w R_g(w)$$

Note 24. Below we show how SGD can be implemented in the learning Problem 1.

Note 25. Assume there is available a finite dataset $\mathcal{S}_n = \{z_i; i = 1, \dots, n\}$ of size n which consists of independent realizations z_i from the data generating distribution g ; $z_i \stackrel{\text{ind}}{\sim} g$. Batch sampling (Algorithm 26) is an implementation of the SGD (Algorithm 6) in the learning Problem 1.

Algorithm 26. batch Stochastic Gradient Descent with learning rate $\eta_t > 0$, and batch size m , for Problem 1.

For $t = 1, 2, 3, \dots$ iterate:

- (1) get a random sub-sample $\{\tilde{z}_j^{(t)}; j = 1, \dots, m\}$ of size m with or without replacement from the complete data-set \mathcal{S}_n .
- (2) compute

$$(5.2) \quad w^{(t+1)} = w^{(t)} - \eta_t v_t,$$

where $v_t = \frac{1}{m} \sum_{j=1}^m \tilde{v}_{t,j}$ and $\tilde{v}_{t,j} \in \partial_w \ell(w^{(t)}, \tilde{z}_j^{(t)})$.

- (3) terminate if a termination criterion is satisfied

Remark 27. Step 1 can be presented equivalently as

- (1) Randomly generate a set $\mathcal{J}^{(t)} \subseteq \{1, \dots, n\}^m$ of m indices from 1 to n with or without replacement, and set a $\tilde{\mathcal{S}}_m = \{z_i; i \in \mathcal{J}^{(t)}\}$.

Hence $v_t = \frac{1}{m} \sum_{j \in \mathcal{J}^{(t)}} v_{t,j}$ where $\tilde{v}_{t,j} \in \partial_w \ell(w^{(t)}, z_j)$.

Remark 28. A projection step, such as (3.2) can be added right after step 2 in Algorithm 26, if needed.

Remark 29. If it is possible to sample anytime fresh examples z_i directly from the data generation model g instead of just having access to only a given finite dataset of examples \mathcal{S}_n , then step 1 in Algorithm 26 can become

(1) sample $\tilde{z}_j^{(t)} \sim g(\cdot)$ for $j = 1, \dots, m$.

Definition 30. Online Stochastic gradient descent is the special case of Algorithm 26 using subsamples of size one ($m = 1$), namely, only one example is randomly chosen in Step 1 of Algorithm 26.

Remark 31. In theory, using larger batch size m has the benefit that reduces the variance $\text{Var}(v_t | w^{(t)})$ of v_t at iteration t due to averaging effect, stabilizes the SGD algorithm, and reduces the error bound (4.1); see Remark 18.

Remark 32. In practice, for a given fixed computational time, using smaller batch size m has the benefit that the algorithm iterates faster as each iteration processes less number of examples. E.g., consider the extreme cases GD vs online SGD in a scenario with big-data: if the dataset consists of several replications of the same values, GD (using all the data) has to process the same information multiple times, while the online SGD (using only one example at a time) would avoid this issue.

Remark 33. In practice, for a given fixed computational time, it is possible for a batch SGD with smaller batch size m to present better generalization properties (wrt the theoretical assumptions) than those with larger m (GD is included). It is observed for the former to be often less prone to getting stuck in shallow local minima because of the additional amount of “noise” E.g., consider the extreme cases GD vs online SGD in a scenario with non-convex risk function (e.g. our theoretical assumptions as violated): if the Risk function presents local minima, considering less examples randomly chosen each time may cause fluctuations in the gradient that allow the chain to accidentally jump/escape to an area with a lower minimum.

6. STOCHASTIC VARIANCE REDUCED GRADIENT (SVRG)

Remark 34. Recall the upper bound of the error (4.1) in SGD depends on the variance of the stochastic gradient, as shown in Remark 4.5. Hence the algorithm may be improved by reducing the variance of each element of v_t .

Remark 35. Control variate is a general way to perform variance reduction. Let random variables $v \in \mathbb{R}$, and $y \in \mathbb{R}$. Let $z = v + c(y - \text{E}(y))$ for some constant $c \in \mathbb{R}$. It is $\text{E}_c(z) = \text{E}(v)$ and

$$\text{Var}_c(z) = \text{Var}(v) + c^2 \text{Var}(y) + 2c\text{Cov}(x, y)$$

which is minimized for

$$c^* = -\frac{\text{Cov}(v, y)}{\text{Var}(y)}$$

hence

$$\text{Var}_{c^*}(z) = \text{Var}(v) - \frac{(\text{Cov}(v, y))^2}{\text{Var}(y)}$$

Note 36. For simplicity, SVRG is presented in the case where $\kappa = 1$ and the loss function $\ell(w, z)$ is differentiable wrt w hence its gradient exists. However it is applicable to the more general cases introduced.

Remark 37. Every m iterations, SVRG keeps a snapshot \tilde{w} , and computes the gradient using all data i.e. $\frac{1}{n} \sum_{i=1}^n \ell(\tilde{w}, z_i)$. At each iteration t , the update is

$$w^{(t+1)} = w^{(t)} - \eta_t \left[\underbrace{\nabla \ell(w^{(t)}, \tilde{z}^{(t)})}_{=v} - \underbrace{\frac{1}{n} \sum_{i=1}^n \nabla \ell(\tilde{w}, z_i)}_{=c} \underbrace{\left(\nabla \ell(\tilde{w}, \tilde{z}^{(t)}) - \frac{1}{n} \sum_{i=1}^n \nabla \ell(\tilde{w}, z_i) \right)}_y \right]$$

given that a random $\tilde{z}^{(t)}$ has been collected from the sample. The symbols below the brackets are given with reference to Remark 35.

Algorithm 38. *Stochastic Variance Reduced Gradient with learning rate $\eta_t > 0$ for Problem 1.*

For $t = 1, 2, 3, \dots$ iterate:

- (1) randomly get an example $\tilde{z}^{(t)}$ from S_n .
- (2) compute

$$(6.1) \quad w^{(t+1)} = w^{(t)} - \eta_t \left[\nabla \ell(w^{(t)}, \tilde{z}^{(t)}) - \nabla \ell(\tilde{w}, \tilde{z}^{(t)}) + \frac{1}{n} \sum_{i=1}^n \nabla \ell(\tilde{w}, z_i) \right]$$

- (3) if $\text{modulo}(t, \kappa) = 0$, set $\tilde{w} = w^{(t)}$
- (4) if a termination criterion is satisfied STOP

Remark 39. Iterations of SVRG are computationally faster than those of full GD, but SVRG can still match the theoretical convergence rate of GD.

Remark 40. How often we get snapshots, κ , in Algorithm 38 is specified by the researcher. The **smaller** the κ , the more frequent snapshots, and the more correlated the baseline y will be with the objective x and hence the bigger the performance improvement; however the iterations will be slower.

7. PRECONDITIONED SGD; THE ADAGRAD ALGORITHM

Remark 41. All SGD (introduced earlier) are first order methods in the sense they consider only the gradient. Their advantage is each iteration is fast. The disadvantage is that they ignore the curvature of the space and hence can be slower to converge in cases the curvature changes eg. among dimensions of w .

Remark 42. To address this, SGD (Algorithm 6) can be modified in the update step as in Algorithm 43 by using a preconditioner P_t that accounts for the curvature (or geometry in general of f).

Algorithm 43. Preconditioned Stochastic Gradient Descent with learning rate $\eta_t > 0$, and preconditioner P_t for the solution of the minimization problem (2.1)

For $t = 1, 2, 3, \dots$ iterate:

(1) compute

$$(7.1) \quad w^{(t+1)} = w^{(t)} - \eta_t P_t v_t,$$

where v_t is a random vector such that $E(v_t|w^{(t)}) \in \partial f(w^{(t)})$, and P_t a preconditioner

(2) terminate if a termination criterion is satisfied, e.g.

If $t \geq T$ then STOP

Remark 44. A natural choice of P_t can be $P_t := [H_t + \epsilon I_d]^{-1}$, where H_t is the Hessian matrix $[H_t]_{i,j} = \frac{\partial^2}{\partial w_i \partial w_j} f(w) \Big|_{w=w^{(t)}}$ (ie. the gradient of the gradient's elements), and ϵ is a tiny $\epsilon > 0$ to mitigate machine error when Hessian elements are close to zero.

Remark 45. If the preconditioner P_t is set to be the inverse of the full Hessian, it may be too expensive to perform matrix operations in (7.1) with the full Hessian, and such operations can be too unstable/inaccurate due to the random error induced by the stochasticity of the gradient.

7.1. Adaptive Stochastic Gradient Decent (AdaGrad).

Remark 46. AdaGrad dynamically incorporates knowledge of the geometry of function $f(\cdot)$ (to be minimized) in earlier iterations to perform more informative gradient-based learning.

Remark 47. AdaGrad aims to perform larger updates (i.e. high learning rates) for those dimensions of w that are related to infrequent features (largest partial derivative) and smaller updates (i.e. low learning rates) for frequent ones (smaller partial derivative).

Remark 48. Hence, this strategy often improves convergence performance over standard stochastic gradient descent in settings where data is sparse and sparse features w 's are more informative.

Definition 49. Adaptive Stochastic Gradient Decent (AdaGrad) can be presented in terms of preconditioned SGD (Algorithm 43) with preconditioner $P_t = [\mathbf{I}_d \text{diag}(G_t) \mathbf{I}_d + \epsilon \mathbf{I}_d]^{-1/2}$ in (7.1)

$$(7.2) \quad w^{(t+1)} = w^{(t)} - \eta_t [\mathbf{I}_d \text{diag}(G_t) \mathbf{I}_d + \epsilon \mathbf{I}_d]^{-1/2} v_t,$$

where $\text{diag}(A) := (A_{1,1}, A_{2,2}, \dots, A_{d,d})^\top$ is the vector of the diagonal of a matrix A , $G_t = \sum_{\tau=1}^t v_\tau^\top v_\tau$ is the sum of the outer products of the gradients $\{v_\tau; \tau \leq t\}$ up to the state t , and $\epsilon > 0$ is a tiny value (eg, 10^{-6}) set for computational stability in case the gradient becomes too close to zero.

Remark 50. AdaGrad algorithm individually adapts the learning rate of each dimension of w_t by scaling them inversely proportional to the square root of the sum of all the past squared values of the gradient $\{v_\tau; \tau \leq t\}$. This is because

$$(7.3) \quad [G_t]_{j,j} = \sum_{\tau=1}^t (v_{\tau,j})^2$$

where j denotes the j -th dimension of w . Hence (7.2) implies that the j -th dimension of w is updated as

$$w_j^{(t+1)} = w_j^{(t)} - \eta_t \frac{1}{\sqrt{[G_t]_{j,j} + \epsilon}} v_{t,j}.$$

Remark 51. The accumulation of positive terms in (7.3) makes the sum keep growing during training and causes the learning rate to shrink and becoming infinitesimally small. This offers an automatic way to choose a decreasing learning rate simplifying setting the learning rate; however it may result in a premature and excessive decrease in the effective learning rate. This can be mitigated by still considering in (7.2) a (user specified rate) $\eta_t \geq 0$ and tuning it properly via pilot runs.

8. EXAMPLE¹

Note 52. We continue Example in Section 5 in Handout 2. Recall, we considered a hypothesis space \mathcal{H} of linear functions $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ with $h(w) = w_1 + w_2 x$, $w = (w_1, w_2)^\top$, and $\ell(w, z = (x, y)^\top) = (y_i - w_1 - w_2 x)^2$. Here we consider a big dataset $\mathcal{S}_n = \{z_1, \dots, z_n\}$ with $n = 10^6$ examples.

Example 53. The batch SGD algorithm (Algorithm 26) with learning rate η_t and batch size $m = 10$ is

- For $t = 1, 2, 3, \dots$ iterate:
 - (1) Randomly generate a set $\mathcal{J}^{(t)}$ by drawing $m = 10$ numbers from $\{1, \dots, n = 10^6\}$, and set $\tilde{\mathcal{S}}_m = \{z_i; i \in \mathcal{J}^{(t)}\}$
 - (2) compute

$$w^{(t+1)} = w^{(t)} - \eta_t v_t,$$

where

$$v_t = \begin{pmatrix} 2w_1^{(t)} + 2w_2^{(t)} \frac{1}{m} \sum_{j \in \mathcal{J}^{(t)}} x_j - 2 \frac{1}{m} \sum_{j \in \mathcal{J}^{(t)}} y_j \\ 2w_1^{(t)} \bar{x} + 2w_2^{(t)} \frac{1}{m} \sum_{j \in \mathcal{J}^{(t)}} x_j^2 - 2 \sum_{j \in \mathcal{J}^{(t)}} y_j x_j \end{pmatrix},$$

- (3) if $t \geq T = 1000$ STOP

because

$$v_t = \frac{1}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla \ell(w^{(t)}, \tilde{z}^{(t)}) = \dots = \begin{pmatrix} 2w_1^{(t)} + 2w_2^{(t)} \frac{1}{m} \sum_{j \in \mathcal{J}^{(t)}} x_j - 2 \frac{1}{m} \sum_{j \in \mathcal{J}^{(t)}} y_j \\ 2w_1^{(t)} \bar{x} + 2w_2^{(t)} \frac{1}{m} \sum_{j \in \mathcal{J}^{(t)}} x_j^2 - 2 \sum_{j \in \mathcal{J}^{(t)}} y_j x_j \end{pmatrix}$$

In Figures 8.1a & 8.1d, we observe that increasing the batch size has improved the convergence however this is not a panacea. Also it had reduced the oscillations of chain $\{w^{(t)}\}$.

Example 54. The SVRG with learning rate $\eta_t > 0$ and batch size $m = 1$ is

- For $t = 1, 2, 3, \dots$ iterate:
 - (1) randomly generate a set $\mathcal{J}^{(t)} = \{j^*\}$ by drawing one number j^* from $\{1, \dots, n = 10^6\}$, and set $\tilde{\mathcal{S}}_1 = \{z_{j^*}\}$

¹Code can be found in https://github.com/georgios-stats/Machine_Learning_and_Neural_Networks_III_Epiphanys_2023/tree/main/Lecture_handouts/code/03.Stochastic_gradient_descent

(2) compute

$$(8.1) \quad w^{(t+1)} = \begin{bmatrix} w_1^{(t)} \\ w_2^{(t)} \end{bmatrix} - \eta_t \left(\begin{bmatrix} 2(w_1^{(t)} - \tilde{w}_1^{(t)}) + 2w_2^{(t)}x_{j^*} \\ 2(w_2^{(t)} - \tilde{w}_2^{(t)})\bar{x} + 2(w_2^{(t)}(x_{j^*})^2 - \tilde{w}_2^{(t)}(x_{j^*})^2) \end{bmatrix} + \nabla \hat{R}_{\mathcal{D}}(\tilde{w}) \right)$$

(3) if modulo $(t, \kappa) = 0$,

(a) set $\tilde{w} = w^{(t)}$

(b) compute

$$(8.2) \quad \frac{1}{n} \sum_{i=1}^n \ell(\tilde{w}, z_i) = \begin{pmatrix} 2\tilde{w}_1^{(t)} + 2\tilde{w}_2^{(t)}\bar{x} - 2\bar{y} \\ 2\tilde{w}_1^{(t)}\bar{x} + 2\tilde{w}_2^{(t)}\bar{x}^2 - 2\bar{y}^\top x \end{pmatrix} = \nabla \hat{R}_{\mathcal{D}}(\tilde{w})$$

(4) if a termination criterion is satisfied STOP

Because (8.2) is actually the gradient of the Risk function at \tilde{w} and

$$\nabla \ell(w^{(t)}, z_{j^*}) - \nabla \ell(\tilde{w}, z_{j^*}) = \begin{pmatrix} 2(w_1^{(t)} - \tilde{w}_1^{(t)}) + 2w_2^{(t)}x_{j^*} \\ 2(w_2^{(t)} - \tilde{w}_2^{(t)})\bar{x} + 2(w_2^{(t)}(x_{j^*})^2 - \tilde{w}_2^{(t)}(x_{j^*})^2) \end{pmatrix}$$

In Figure 8.1c we observe the frequency of the snapshots has improved the convergence; however this is not a panacea as seen in Figure 8.1f.

Example 55. The AdaGrad with $\eta_t = 1$, $\epsilon = 10^{-6}$, and batch size $m = 1$ is

- Set $G_0 = (0, 0)^\top$.
- For $t = 1, 2, 3, \dots$ iterate:
 - (1) randomly generate a set $\mathcal{J}^{(t)} = \{j^*\}$ by drawing one number from $\{1, \dots, n = 10^6\}$, and set $\tilde{\mathcal{S}}_1 = \{z_{j^*}\}$
 - (2) compute

$$\begin{aligned} v_t &= \begin{pmatrix} 2w_1^{(t)} + 2w_2^{(t)}x_{j^*} - 2y_{j^*} \\ 2w_1^{(t)}\bar{x} + 2w_2^{(t)}x_{j^*}^2 - 2y_{j^*}x_{j^*} \end{pmatrix} \\ G_t &= G_{t-1} + \begin{pmatrix} v_{t,1}^2 \\ v_{t,2}^2 \end{pmatrix} \\ w^{(t+1)} &= \begin{pmatrix} w_1^{(t)} - \frac{\eta_t}{\sqrt{G_{t,1} + \epsilon}} v_{t,1} \\ w_2^{(t)} - \frac{\eta_t}{\sqrt{G_{t,2} + \epsilon}} v_{t,2} \end{pmatrix}, \end{aligned}$$

(3) if $t \geq T = 1000$ STOP

because

$$v_t = \nabla \ell(w^{(t)}, z_{j^*}) = \dots = \begin{pmatrix} 2w_1^{(t)} + 2w_2^{(t)}x_{j^*} - 2y_{j^*} \\ 2w_1^{(t)}\bar{x} + 2w_2^{(t)}x_{j^*}^2 - 2y_{j^*}x_{j^*} \end{pmatrix}$$

In Figures 8.1g & 8.1h, we see that AdaGrad with $\eta = 1$ works (I did not try to tune it), however to make vanilla SGD to work I have to tune $\eta = 0.03$ otherwise for $\eta = 1.0$ it did not work.

Example 56. Consider a (rather naive) loss function $\ell(w, z = (x, y)) = -\cos(0.5(y - w_1 - w_2x))$, a hypothesis class $\mathcal{H} = \{w \in \mathbb{R}^2 : \|w\| \leq 1.5\}$, and assume that inputs x in dataset \mathcal{D} are such that

$x \in [-1, 1]$. Note that $-\cos(\cdot)$ is convex in $[-1.5, 1.5]$ and non-convex in \mathbb{R} . Consider learning rate $\eta_t = 50/t$ reducing to zero, and seed $w^{(0)} = (1.5, 1.5)$. An unconstrained SGD may produce a minimizer/solution outside \mathcal{H} because η_t is too large at the first few iterations. We can design the online SGD (batch size $m = 1$) with projection to \mathcal{H} as

- For $t = 1, 2, 3, \dots$ iterate:

(1) randomly generate a set $\mathcal{J}^{(t)} = \{j^*\}$ by drawing one number from $\{1, \dots, n = 10^6\}$, and

set $\tilde{\mathcal{S}}_1 = \{z_{j^*}\}$

(2) compute

$$w^{(t+1/2)} = w^{(t)} - \frac{50}{t} \begin{pmatrix} \sin(y_{j^*} - w_1^{(t)} - w_2^{(t)} x_{j^*}) \\ \sin(y_{j^*} - w_1^{(t)} - w_2^{(t)} x_{j^*}) x_{j^*} \end{pmatrix}$$

$$w^{(t+1/2)} = \arg \min_{\forall w: \|w\| \leq 1.5} (\|w - w^{(t+1/2)}\|)$$

(3) if $t \geq T = 1000$ STOP

because

$$v_t = \nabla \ell(w^{(t)}, z_{j^*}) = \dots = \begin{pmatrix} \sin(y_{j^*} - w_1^{(t)} - w_2^{(t)} x_{j^*}) \\ \sin(y_{j^*} - w_1^{(t)} - w_2^{(t)} x_{j^*}) x_{j^*} \end{pmatrix}$$

In Figures 8.1b & 8.1e, we observe that the SGD got trapped outside \mathcal{H} due to the unreasonably large learning rate at the beginning of the iterations, while the SGD with projection step managed to stay in \mathcal{H} and converge.

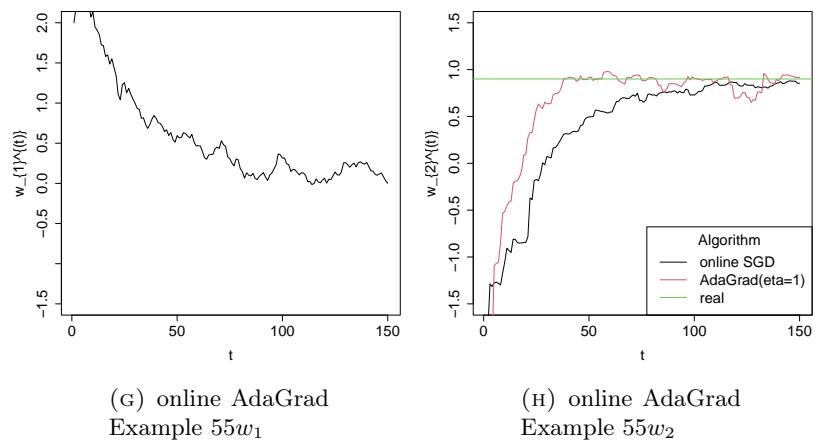
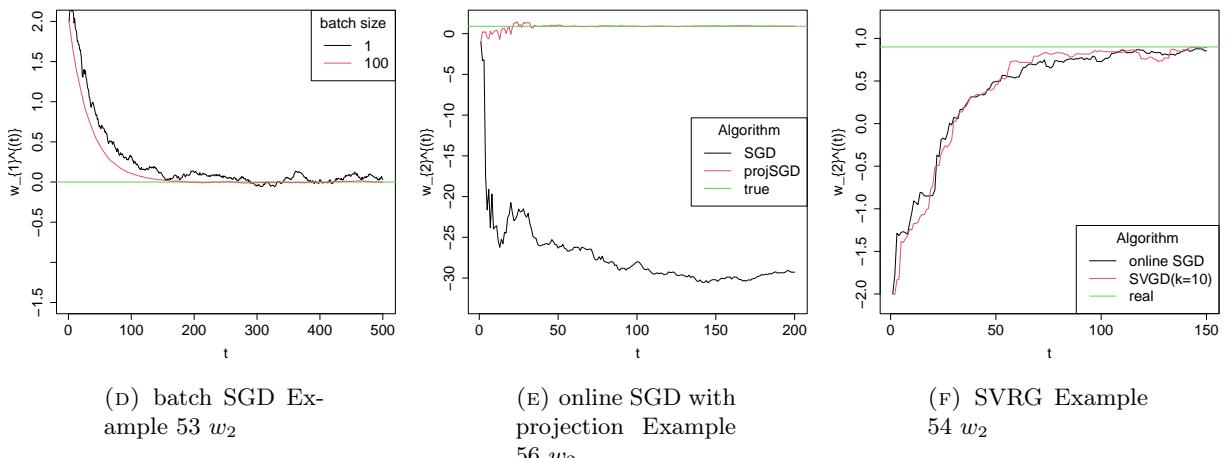
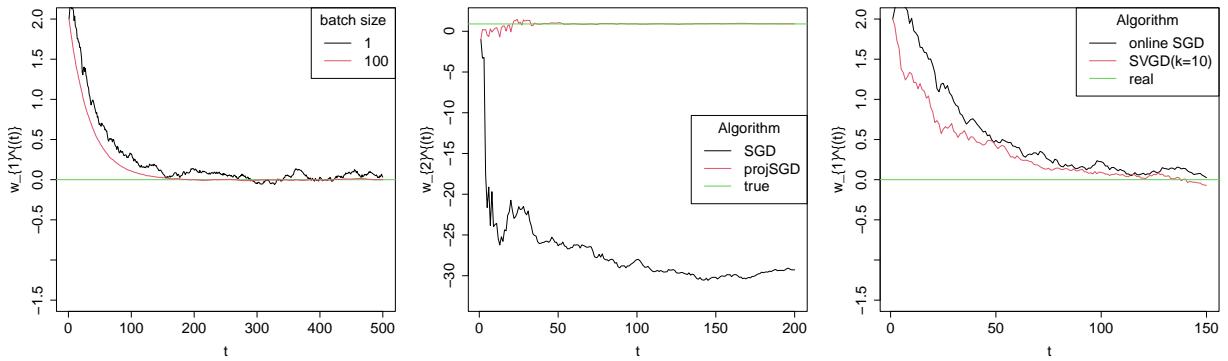


FIGURE 8.1. Simulations of the Examples

Handout 4: Bayesian Learning via Stochastic gradient and Stochastic gradient Langevin dynamics

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce the Bayesian Learning, and Stochastic gradient Langevin dynamics (description, heuristics, and implementation). Stochastic gradient descent will be implemented in the Bayesian learning too.

Reading list & references:

- Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 681-688).
- Bottou, L. (2012). Stochastic gradient descent tricks. In Neural networks: Tricks of the trade (pp. 421-436). Springer, Berlin, Heidelberg.
- Sato, I., & Nakagawa, H. (2014). Approximation analysis of stochastic gradient Langevin dynamics by using Fokker-Planck equation and Ito process. In International Conference on Machine Learning (pp. 982-990). PMLR.
- Teh, Y. W., Thiery, A. H., & Vollmer, S. J. (2016). Consistency and fluctuations for stochastic gradient Langevin dynamics. Journal of Machine Learning Research, 17.
- Vollmer, S. J., Zygalakis, K. C., & Teh, Y. W. (2016). Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics. The Journal of Machine Learning Research, 17(1), 5504-5548. ->further reading for theory
- Nemeth, C., & Fearnhead, P. (2021). Stochastic gradient Markov chain Monte Carlo. Journal of the American Statistical Association, 116(533), 433-450.

1. BAYESIAN LEARNING AND MOTIVATIONS

Remark 1. Bayesian methods are appealing in their ability to capture uncertainty in learned parameters and avoid overfitting. Arguably with large datasets there will be little overfitting. Alternatively, as we have access to larger datasets and more computational resources, we become interested in building more complex models (eg from a logistic regression to a deep neural network), so that there will always be a need to quantify the amount of parameter uncertainty.

Note 2. Consider a Bayesian statistical model with sampling distribution (statistical model) $f(z|w)$ labeled by an unknown parameter $w \in \Theta \subseteq \mathbb{R}^d$ that follows a prior distribution $f(w)$. Assume a dataset $\mathcal{S}_n = \{z_i; i = 1, \dots, n\}$ of size n containing independently drawn examples. Let $L_n(w) := f(z_{1:n}|w)$ denote the likelihood of the observables $\{z_i \in \mathcal{Z}\}_{i=1}^n$ give parameter w . The Bayesian

model is denoted as

$$(1.1) \quad \begin{cases} z_{1:n}|w & \sim f(z_{1:n}|w) \\ w & \sim f(w) \end{cases}$$

Remark 3. Bayesian learning (inference) relies on the posterior distribution density

$$(1.2) \quad f(w|z_{1:n}) = \frac{L_n(w)f(w)}{\int L_n(w)f(w)dw}$$

which quantifies the researcher's belief (or uncertainty) about the unknown parameter w learned given examples $\{z_i \in \mathcal{Z}\}_{i=1}^n$. It is often intractable; hence there is often a need to numerically compute it. (Section 3)

Remark 4. Point estimation of a function h of w is often performed via computation of the posterior expectation w given the examples in \mathcal{S}_n

$$(1.3) \quad E_f(h(w)|z_{1:n}) = \int h(w)f(w|z_{1:n})dw$$

It is often intractable; hence there is often a need to numerically compute it. (Section 3)

Remark 5. Point estimation of w can also be performed via maximum a-posteriori (MAP) estimator w^* of w that is the maximizer w^* of (1.2) i.e.

$$(1.4) \quad w^* = \arg \max_{\forall w \in \Theta} (f(w|z_{1:n}))$$

$$(1.5) \quad = \arg \min_{\forall w \in \Theta} \left(\underbrace{-\log(L_n(w))}_{(I)} - \underbrace{\log(f(w))}_{(II)} \right)$$

Note that (I) may be interpreted as an empirical risk function, and (II) can be interpreted as a shrinkage term in terms of shrinkage methods (like LASSO, Ridge). It is often intractable; hence there is often a need to numerically compute it. (Section 2)

Note 6. To describe the learning algorithms Gradient Descent (GD), Stochastic Gradient Descent (SGD), and Stochastic Gradient Langevin Dynamics (SGLD), we consider that the examples (data) in (1.1) are independent realizations from the sampling distribution i.e. $z_i|w \sim f(\cdot|w)$ for $i = 1, \dots, n$ and that w is continuous (not discrete).

Note 7. In what follows, we first present the implementation of GD and SGD addressing MAP learning, and then we introduce the implementation of SGLD addressing posterior density and expectation learning.

2. MAXIMUM A POSTERIORI (MAP) LEARNING VIA GD AND SGD

Problem 8. Given the Bayesian model (1.1), and rearranging (1.4), MAP estimate w^* of w can be computed as

$$(2.1) \quad w^* = \arg \min_{\forall w \in \Theta} (-\log(L_n(w)) - f(w)) = \arg \min_{\forall w \in \Theta} \left(-\sum_{i=1}^n \log(f(w|z_i)) - \log(f(w)) \right)$$

Remark 9. GD is particularly suitable to solve (2.1) when w has high dimensionality.

Algorithm 10. Gradient descent (Algorithm 1 Handout 2) with learning rate $\eta_t \geq 0$ can be used to solve (2.1) by using the update rule as

For $t = 1, 2, 3, \dots$ iterate:

(1) Compute

$$(2.2) \quad w^{(t+1)} = w^{(t)} + \eta_t \left(\sum_{i=1}^n \nabla_w \log \left(f \left(\mathbf{z}_j | \mathbf{w}^{(t)} \right) \right) + \nabla_w \log \left(f \left(w^{(t)} \right) \right) \right)$$

Remark 11. The implementation of other GD variants (eg (3.2) in Handout 2) is straightforward, based on Algorithm 10.

Remark 12. SGD is particularly suitable to solve (2.1) when w has high dimensionality, and in big-data problems since the repetitive computation of the sum in (2.2) is prohibitively expensive. Yet consider the benefits of SGD against GD as discussed in Remarks 32 and 33 of Handout 3.

Algorithm 13. Batch Stochastic Gradient Descent (Algorithm 26 in Handout 3) with learning rate $\eta_t \geq 0$ and batch size m can be used to solve (2.1) by using the update rule as

For $t = 1, 2, 3, \dots$ iterate:

- (1) generate a random set $\mathcal{J}^{(t)} \subseteq \{1, \dots, n\}^m$ of m indices from 1 to n with or without replacement, and set a $\tilde{\mathcal{S}}_m = \{z_i; i \in \mathcal{J}^{(t)}\}$.
- (2) compute

$$(2.3) \quad w^{(t+1)} = w^{(t)} + \eta_t \left(\frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log \left(f \left(\mathbf{z}_j | \mathbf{w}^{(t)} \right) \right) + \nabla_w \log \left(f \left(w^{(t)} \right) \right) \right)$$

Remark 14. Recursion (2.3) is justified in terms of SGD theory as

$$(2.4) \quad \mathbb{E}_{\mathcal{J}^{(t)} \sim \text{simple-random-sampling}} \left(\frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log \left(f \left(z_j | w^{(t)} \right) \right) \right) = \sum_{i=1}^n \nabla_w \log \left(f \left(z_i | w^{(t)} \right) \right)$$

Remark 15. The implementation of the SGD variants (Algorithms 26, 38, 43, 49 in Handout 3)) is straightforward based on Algorithm 13 and (2.4).

3. FULLY BAYESIAN LEARNING VIA SGLD

Problem 16. Fully Bayesian learning, computationally, is the problem of recovering the posterior distribution $f(w|z_{1:n})$ of w given $z_{1:n}$ that admits density (1.2). For a given Bayesian model (1.1), the Bayesian estimator of $h := h(w)$ can be computed as the posterior expectation of w given the data \mathcal{S}_n

$$(1.3) \quad \mathbb{E}_f(h(w)|z_{1:n}) = \int h(w) f(w|z_{1:n}) dw$$

Remark 17. Monte Carlo integration aims at approximating (1.3), by using Central Limit Theorem or Law of Large Numbers arguments as $\hat{h} \approx E_f(h(w) | z_{1:n})$ where

$$(3.1) \quad \hat{h} = \frac{1}{T} \sum_{t=1}^T h(w^{(t)})$$

where $\{w^{(t)}\}$ are T simulations drawn (approximately) from the posterior distribution 1.2. This theory is subject to conditions we skip.

Remark 18. Stochastic gradient Langevin dynamics (SGLD) algorithm is able to approximately produce samples from the posterior distribution (1.2) of parameters w given the available data $z_{1:n}$. That allows to recover the whole posterior distribution (1.2) (hence account for the uncertainty in parameters) and approximate posterior expectations (1.3) by averaging out (3.1) according to the Monte Carlo integration (Remark 17).

Remark 19. Stochastic gradient Langevin dynamics (SGLD) algorithm is able to generate a sample approximately distributed according to the posterior distribution (1.2). That allows to recover the whole posterior distribution (hence account for uncertainty in parameters) and approximate posterior expectations based on the Monte Carlo integration (Remark 17).

Note 20. SGLD relies on injecting the ‘right’ amount of noise to a standard stochastic gradient optimization recursion (2.2), such that, as the stepsize η_t properly reduces, the produced chain $\{w^{(t+1)}\}$ converges to samples that could have been drawn from the true posterior distribution.

Algorithm 21. *Stochastic Gradient Langevin Dynamics (SGLD) with learning rate $\eta_t > 0$, batch size m , and temperature $\tau > 0$ is*

-
- (1) *Generate a random set $\mathcal{J}^{(t)} \subseteq \{1, \dots, n\}^m$ of m indices from 1 to n with or without replacement, and set a $\tilde{\mathcal{S}}_m = \{z_i; i \in \mathcal{J}^{(t)}\}$.*
 - (2) *Compute*

$$(3.2) \quad w^{(t+1)} = w^{(t)} + \eta_t \left(\frac{n}{m} \sum_{i \in \mathcal{J}^{(t)}} \nabla \log f(z_i | w) + \nabla \log f(w) \right) + \sqrt{\eta_t} \sqrt{\tau} \epsilon_t$$

where $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, 1)$.

- (3) *Terminate if a termination criterion is satisfied; E.g., $t \leq T_{max}$ for a prespecified $T_{max} > 0$.*

Remark 22. The first few iterations from Algorithm 21 involve values generated at the beginning of the running algorithm while the chain have not yet converged to (or reached) an area of substantial posterior mass. Hence they are discarded from the output of the SGLD. These values are called burn-in.

Remark 23. The output of SGLD (Algorithm 21) $\{w^{(t)}\}$ includes the generated values of w produced during the last few iterations of the running algorithm (aka the end tail of the generated chain).

Remark 24. SGLD (Algorithm 21) generates as output a random chain $\{w^{(t)}\}$ that is approximately distributed according to a distribution with density such as

$$(3.3) \quad f_\tau(w|z_{1:n}) \propto \exp\left(\frac{1}{\tau} \prod_{i=1}^n f(z_i|w) f(w)\right)$$

$$(3.4) \quad \propto \exp\left(\frac{1}{\tau} L_n(w) f(w)\right)$$

under regularity conditions. Conditions 25 on the learning rate are rather inevitable and should be satisfied.

Condition 25. Regarding the learning rate (or gain) $\{\eta_t\}$ should satisfy conditions

- (1) $\eta_t \geq 0$,
- (2) $\sum_{t=1}^{\infty} \eta_t = \infty$
- (3) $\sum_{t=1}^{\infty} \eta_t^2 < \infty$

Remark 26. The temperature parameter $\tau > 0$ is user specified and aims at controlling (eg; inflating) the variance of the produced chain for instance with practical purpose to escape from local modes (otherwise energy barriers) in non-convex problems.

Remark 27. SGLD for $\tau = 1$ approximately simulates from the posterior (1.2).

Remark 28. The popular learning rates $\{\eta_t\}$ in Remark 9 in Handout 2 satisfy Condition 25 and hence can be used in SGD too. Once parametrized, η_t can be tuned based on pilot runs using a reasonably small number of data.

Remark 29. (Mathematically speaking) The stochastic chain in (3.2) can be viewed as a discretization of the continuous-time Langevin diffusion described by the stochastic differential equation

$$(3.5) \quad dW(t) = -\nabla_w [-\log f(W(t)|z_{1:n})] dt + \sqrt{2\tau} dB(t), \quad t \geq 0$$

where $\{B(t)\}$ is a standard Brownian motion¹ in \mathbb{R}^d (i.e.). Under suitable assumptions on f , it can be shown that a Gibbs distribution with PDF such as

$$(3.6) \quad f^*(w|z_{1:n}) \propto \exp\left(-\frac{1}{\tau} [-\log f(W(t)|z_{1:n})]\right)$$

is the unique invariant distribution of (3.5), and that the distributions of $W(t)$ converge rapidly to f^* as $t \rightarrow \infty$.

Remark 30. (Heuristically speaking) In the initial phase of running, the stochastic gradient noise will dominate the injected noise ϵ_t and the algorithm will imitate an efficient SGD Algorithm 10 -but this is until η_t or $\nabla \log(L_n(w))$ become small enough. In the later phase of running, the injected noise ϵ_t will dominate the stochastic gradient noise, so the SGLD will imitate a Langevin dynamics for the target distribution (1.2). The aim is for the algorithm to transition smoothly between the

¹A continuous-time stochastic process: (1) $B(0) = 0$; (2) $B(t)$ is almost surely continuous ; (3) $B(t)$ has independent increments ; (4) $B(t) - B(s) \sim N(0, t-s)$ for $0 \leq s \leq t$.

two phases. Whether the algorithm is in the stochastic optimization phase or Langevin dynamics phase depends on the variance of the injected noise versus that of the stochastic gradient.

Remark 31. One can argue that, the output of SGLD is also an ‘‘almost’’ minimizer of the empirical risk for large enough t . A draw from the Gibbs distribution (3.6) is approximately a minimizer of (2.1). Also one can show that the SGLD recursion tracks the Langevin diffusion (3.5) in a suitable sense. Hence, both imply that the distributions of $W(t)$ will be close to the Gibbs distribution (3.6) for all sufficiently large t .

Remark 32. To guarantee the algorithm to work it is important for the step sizes η_t to decrease to zero, so that the mixing rate of the algorithm will slow down with increasing number of iterations t . Then, we can keep the step size η_t constant once it has decreased below a critical level.

Remark 33. Expectation (1.3), can be estimated as an arithmetic average

$$(3.7) \quad \widehat{h_T(w)} = \frac{1}{T} \sum_{t=1}^T h(w^{(t)})$$

as $\widehat{h_T(w)} \rightarrow E_f(h(w) | z_{1:n})$ based on LLN arguments.

Remark 34. Another more efficient estimator for the expectation (1.3) is the weighted arithmetic average

$$(3.8) \quad \widehat{h(w)} = \sum_{t=T_0+1}^T \frac{\eta_t}{\sum_{t=T_0+1}^T \eta_t} h(w^{(t)})$$

Because the step size η_t decreases, the mixing rate of the chain $\{w^{(t)}\}$ decreases as well and the simple sample average (3.7) will overemphasize the tail end of the sequence where there is higher correlation among the samples resulting in higher variance in the estimator.

Remark 35. Certain dimensions may have a vastly larger curvature leading to much bigger gradients. In this case a symmetric preconditioning matrix $P_t > 0$ can transform all dimensions to the same scale; this is similar to the SGD case in Handout 3. Hence the update (3.2) becomes

$$(3.9) \quad w^{(t+1)} = w^{(t)} + \eta_t P_t \left(\frac{n}{m} \sum_{i \in J^{(t)}} \nabla \log f(z_i | w) + \nabla \log f(w) \right) + \sqrt{\eta_t} \sqrt{\tau} P_t^{\frac{1}{2}} \epsilon_t$$

where $P_t^{\frac{1}{2}}$ is such that $P_t^{\frac{1}{2}} \left(P_t^{\frac{1}{2}} \right)^{\top} = P_t$.

Remark 36. ‘Exploding gradients’ is the practical phenomenon in which large updates to weights during training can cause a numerical overflow or underflow due to the machine error of the computer. Practical solutions involve, at each iteration t , checking the magnitude of the gradient v_t , (e.g., Euclidean norm $\|v_t\|$), and instantly changing it (e.g., truncating it) if it is gonna result an overflow.

Gradient scaling: involves normalizing the gradient vector such that vector norm (magnitude) equals a defined value. More formally, given any gradient v on an example, gradient

clipping can be used in a standard recursion

$$w^{(t+1)} = w^{(t)} + \eta_t v_t$$

as

$$w^{(t+1)} = w^{(t)} + \eta_t \text{clip}(v_t, c)$$

where

$$\text{clip}(v, c) = v \min\left(1, \frac{c}{\|v\|}\right)$$

and c is a clipping threshold implying that clipping will take place at iteration t if $\|v_t\| > c$.

Gradient clipping: involves forcing the gradient values (element-wise) to a specific minimum or maximum value if the gradient exceeded an expected range.

Beware that unreasonable clipping may introduce significant bias and hence it should not be applied unnecessarily.

4. EXAMPLES ²

We continue the Example 33 in Handout 1, and Example 8 in Handout 2. Consider the Bayesian Normal linear regression model

$$\begin{cases} y_i|\beta, \sigma^2 \sim N(x_i^\top \beta, \sigma^2) & \text{sampling distribution } f(y_i|\beta, \sigma^2) \\ \beta|\sigma^2 \sim N(\mu, \sigma^2 V) & \text{prior } f(\beta|\sigma^2) \\ \sigma^2 \sim IG(\phi, \psi) & \text{prior } f(\sigma^2) \end{cases}$$

and $f(\beta, \sigma^2) = f(\beta|\sigma^2) f(\sigma^2)$, $\beta \in \mathbb{R}^d$, and $\sigma^2 \in \mathbb{R}_+$. Note given densities

$$\begin{aligned} N(x|\mu, \Sigma) &= \left(\frac{1}{2\pi}\right)^d \frac{1}{|\Sigma|} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu)\right) \\ IG(x|a, b) &= \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{b}{x}\right) \mathbf{1}(x \geq 0) \end{aligned}$$

Because SGD (Algorithm 13) and SGLD (Algorithm 21) can handle cases where $w \in \mathbb{R}^d$ in a straightforward manner than what they do when $w = (\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$ which requires an additional projection step; we consider a transformation $w = (\beta, \gamma) \in \mathbb{R}^d \times \mathbb{R}$ with $\gamma = \log(\sigma^2)$. Hence, the Bayesian model becomes

$$\begin{cases} y_i|\beta, \sigma^2 \sim N(x_i^\top \beta, \exp(\gamma)) & \text{sampling distribution } f(y_i|\beta, \gamma) \\ \beta|\sigma^2 \sim N(\mu = 0, \exp(\gamma) V) & \text{prior } f(\beta|\gamma) \\ \gamma \sim f_\gamma(\gamma) & \text{prior } f(\gamma) \end{cases}$$

where $f_\gamma(\gamma)$ is computed according to the method of bijective transformation of random variables; i.e.

$$f_\gamma(\gamma) = IG(\exp(\gamma)|\phi, \psi) \left| \frac{d}{d\gamma} \exp(\gamma) \right| = IG(\exp(\gamma)|\phi, \psi) \exp(\gamma)$$

²Code is available from https://github.com/georgios-stats/Machine_Learning_and_Neural_Networks_III_Epiphanys_2023/tree/main/Lecture_handouts/code/04.Stochastic_gradient_Langevine_dynamics

Then we can compute the required gradients in order to run the SGD, and SGLD with respect to $w = (\beta, \gamma)$ with $\gamma = \log(\sigma^2)$.

$$\begin{aligned}\log(f(z_i = (x_i, y_i) | w)) &= -\frac{1}{2} \log(2\pi) - \frac{1}{2}\gamma - \frac{1}{2}(y_j - x_i^\top \beta)^2 \exp(-\gamma) \\ \log(f(w = (\beta, \gamma))) &= \log(f(\beta|\gamma)) + \log(f(\gamma)) \\ \log(f(\beta|\gamma)) &= -\frac{d}{2} \log(2\pi) - \frac{d}{2}\gamma - \frac{1}{2}|V| - \frac{1}{2}\exp(-\gamma)(\beta - \mu)^\top V^{-1}(\beta - \mu) \\ \log(f(\gamma)) &= \psi \log(\phi) - \log(\Gamma(\phi)) - (\phi + 1)\gamma - \psi \exp(-\gamma) + \gamma\end{aligned}$$

Hence for the log sampling PDF we have

$$\begin{aligned}\nabla_w \log(f(z_i|w)) &= \left(\frac{d}{d\beta} \log(f(z_i|w)), \frac{d}{d\gamma} \log(f(z_i|w)) \right) \\ \frac{d}{d\beta} \log(f(z_i|w)) &= (y_i - x_i^\top \beta) x_i \exp(-\gamma) \\ \frac{d}{d\gamma} \log(f(z_i|w)) &= -\frac{1}{2} + \frac{1}{2}(y_j - x_i^\top \beta)^2 \exp(-\gamma)\end{aligned}$$

...so

$$(4.1) \quad \nabla_w \log(f(z_i|w)) = \begin{pmatrix} (y_i - x_i^\top \beta) x_i \exp(-\gamma) \\ -\frac{1}{2} + \frac{1}{2}(y_j - x_i^\top \beta)^2 \exp(-\gamma) \end{pmatrix}$$

Hence for the log a priori PDF we have

$$\begin{aligned}\nabla_w \log(f(w)) &= \left(\frac{d}{d\beta} \log(f(w)), \frac{d}{d\gamma} \log(f(w)) \right) \\ \frac{d}{d\beta} \log(f(w)) &= -\exp(-\gamma) V^{-1}(\beta - \mu) \\ \frac{d}{d\gamma} \log(f(w)) &= -\frac{d}{2} + \frac{1}{2}\exp(-\gamma)(\beta - \mu)^\top V^{-1}(\beta - \mu) - (\phi + 1) + \psi \exp(-\gamma) + 1\end{aligned}$$

...so

$$(4.2) \quad \nabla_w \log(f(w)) = \begin{pmatrix} -\exp(-\gamma) V^{-1}(\beta - \mu) \\ \frac{d}{2} + \frac{1}{2}\exp(-\gamma)(\beta - \mu)^\top V^{-1}(\beta - \mu) - (\phi + 1) + \psi \exp(-\gamma) + 1 \end{pmatrix}$$

To implement SGD (Algorithm 13) and SGLD (Algorithm 21), we just need to plug in the computed gradients (4.1) and (4.2) for $w = (\beta, \gamma) \in \mathbb{R}^d \times \mathbb{R}$ with $\gamma = \log(\sigma^2)$. After running SGD and SGLD with the computed gradients with respect to $w = (\beta, \gamma) \in \mathbb{R}^d \times \mathbb{R}$ with $\gamma = \log(\sigma^2)$, and obtaining chains $\{w^{(t)} = (\beta^{(t)}, \gamma^{(t)})\}_{t=1}^T$, we can just perform transformation $\{(\sigma^{(t)})^2 = \exp(\gamma^{(t)})\}_{t=1}^T$ if we are interested in learning (β, σ^2) .

We consider $\mu = 0$, $\phi = 1$, $\psi = 1$, and $V = 100I_d$, for our simulations below.

- In Figures 4.1, we ran the SGD for different batch sizes m and compared it against the exact MLE. We observe that SGD trace converges to the exact MLE. The oscillations are due to the stochastic gradient (ie, noise in the gradient).
- In Figures 4.2, we ran the SGLD for different batch sizes m but the same temperature $\tau = 1$ and compared it against the exact posterior densities. We observe that the histograms of

$\{w^{(t)}\}$ produced from SGLD are closer to the curves representing the exact posteriors when the batch size m is bigger. As we said this is not a panacea; if the landscape of the exact posterior density was multimodal (aka not convex but with had several maxima), then the SGLD using smaller batch sizes could have performed better, in the sense that the inflated noise from the stochastic gradient could accidentally make the generated chain to pass the low mass barrier and visit a different mode, unlike the one with larger batch-size and hence smaller variation.

- In Figures 4.3, we ran the SGLD for different temperatures τ but the same batch sizes $m = 100$ and compared it against the exact posterior densities. We observe that increasing the temperature τ may increase the variation of the produced chain. We can use a large τ at the beginning of the run of the algorithm to perform an exploration of the space (this is particularly useful for non-convex/multimodal densities as it allows visiting different modes), and later on we can use a smaller temperature such as $\tau = 1$.

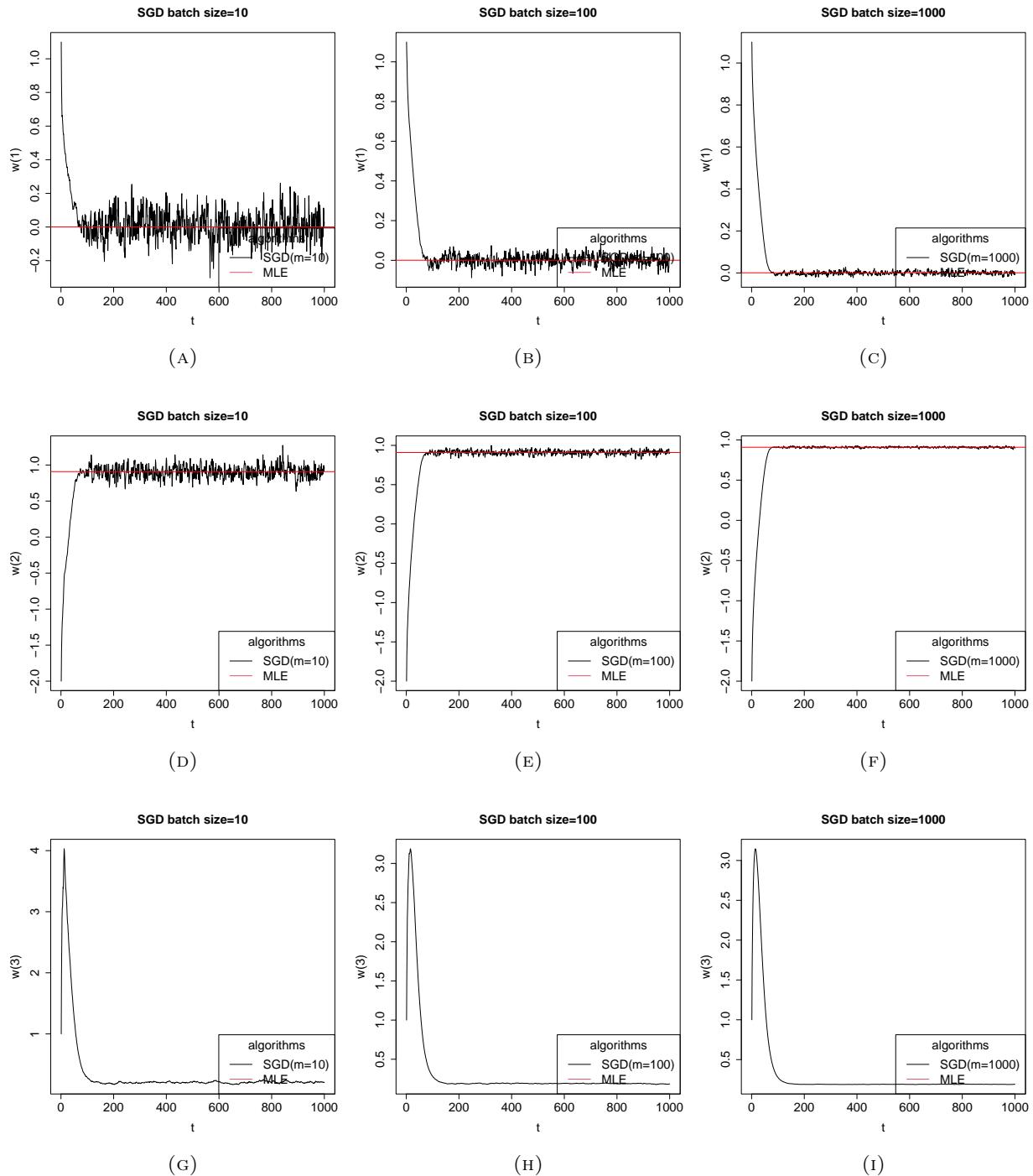


FIGURE 4.1. Bayesian learning via SGD (SGD vs (exact)MLE) Study on batch size m .

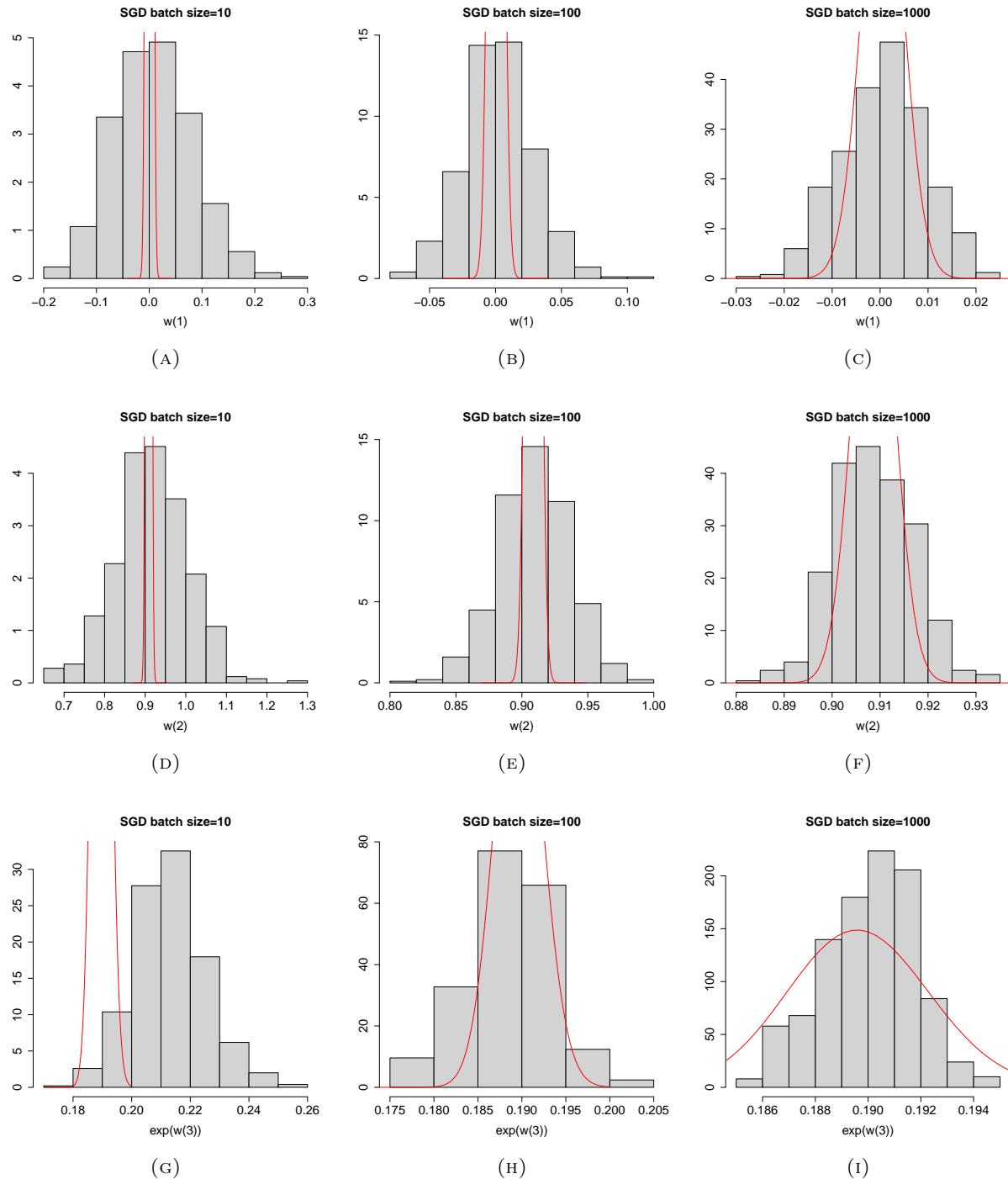


FIGURE 4.2. Bayesian learning: SGLD vs exact posterior (in red). Temperature $\tau = 1$. Study on batch size m

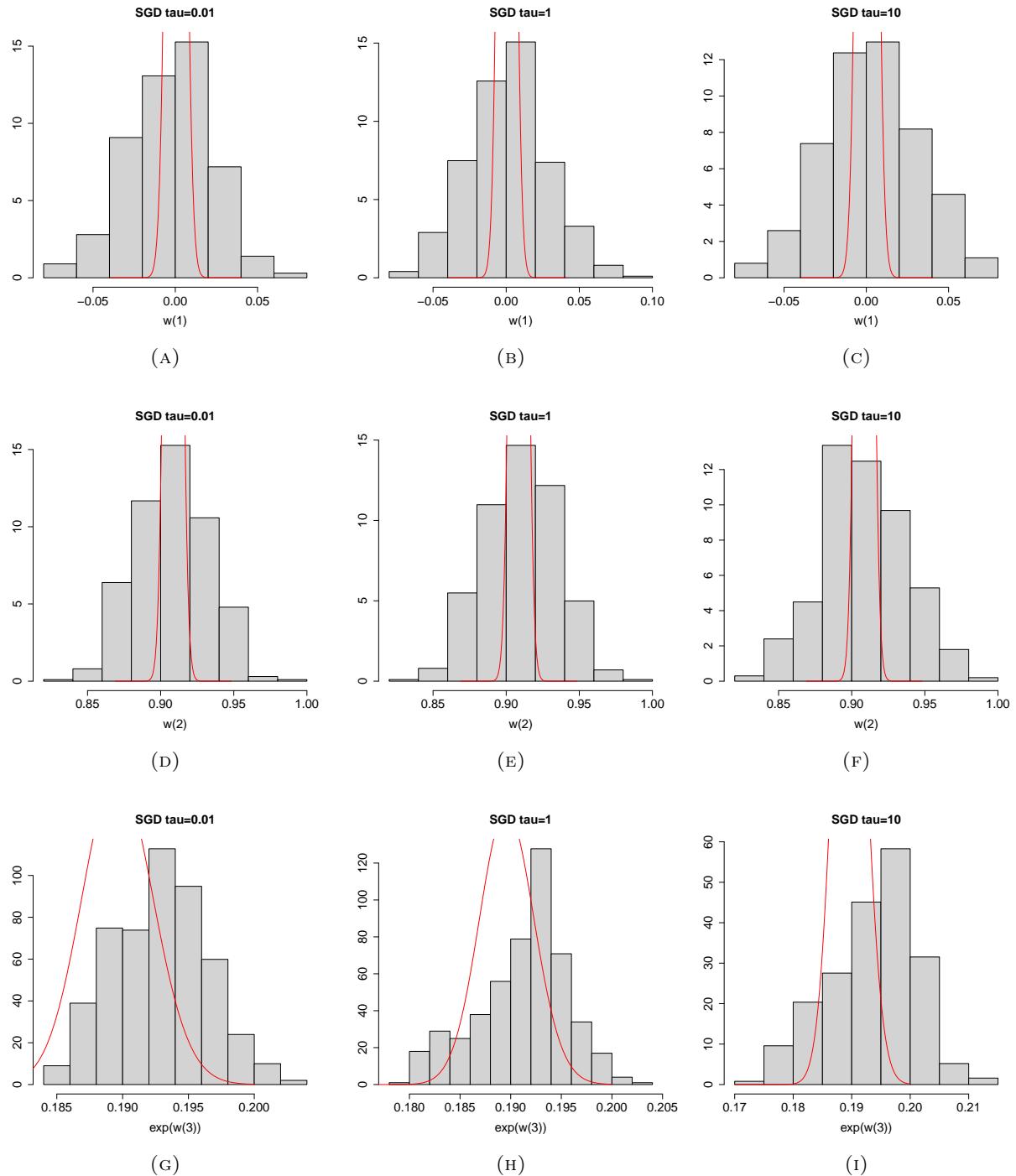


FIGURE 4.3. Bayesian learning: SGLD vs exact posterior (in red). Batch size = 100. Study on τ

Handout 5: Artificial neural networks

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce the Artificial neural network as a model and procedure in classical and Bayesian framework. Motivation, set-up, description, computation, implementation, tricks. We focus on the Feedforward network.

Reading list & references:

- (1) Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
 - Ch. 20 Neural Networks
- (2) Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.
 - Ch. 5 Neural Networks
- (3) Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.
 - Ch. 4 The multi-layer perceptron
- (4) LeCun, Y., Bottou, L., Orr, G. B., & Müller, K. R. (2002). Efficient backprop. In Neural networks: Tricks of the trade (pp. 9-50). Berlin, Heidelberg: Springer Berlin Heidelberg.

1. INTRO AND MOTIVATION ¹

Note 1. Artificial Neural Networks (NN) are statistical models which have mostly been developed from the algorithmic perspective of machine learning. They were originally created as an attempt to model the act of thinking by modeling neurons in a brain. In ML, NN are used as global approximators.

Note 2. The original biological motivation for feed-forward NN stems from McCulloch & Pitts (1943) who published a seminal model of a NN as a binary thresholding device in discrete time, i.e.

$$n_j(t) = 1 \left(\sum_{\forall i \rightarrow j} w_{j,i} n_i(t-1) > \theta_i \right)$$

where the sum is over neuron i connected to neuron j ; $n_j(t)$ is the output of neuron i at time t and $0 < w_{j,i} < 1$ are attenuation weights. Thus the effect is to threshold a weighted sum of the inputs at value θ_i . Perhaps, such a mathematical model involving compositions of interconnected non-linear functions could be able to mimic human's learning mechanism and be implemented in a computing environment (with faster computational abilities) with purpose to discover patterns, make predictions, cluster, classify, etc....

¹In this Section, formulas not needed to be memorized.

Remark 3. Mathematically, NN are rooted in the classical theorem by Kolmogorov stating (informally) that every continuous function $h(\cdot)$ on $[0, 1]^d$ can be written as

$$(1.1) \quad h(x) = \sum_{i=1}^{2d+1} F_i \left(\sum_{j=1}^d G_{i,j}(x_j) \right)$$

where $\{G_{i,j}\}$ and $\{F_i\}$ are continuous functions whose form depends on f . Perhaps, one may speculate that functions $\{G_{i,j}\}$ and $\{F_i\}$ can be approximated by sigmoids or threshold functions of the form $\sigma(w^\top x)$ allowing the number of the tunable coefficients w to be high enough such that they can represent any function -hence the property of NN as global approximators.

Example 4. Consider a regression problem with predictive rule $h : \mathbb{R}^d \rightarrow \mathbb{R}^q$, suitable for cases where the examples (data) consist of input $x \in \mathbb{R}^d$, and output targets $y \in \mathbb{R}^q$. A 2 layer artificial neural network is

$$h_k(x) = \sigma_{(2)} \left(w_{(2),k,0} + \sum_{\forall j} w_{(2),k,j} \sigma_{(1)} \left(w_{(1),j,0} + \sum_{\forall i} w_{(1),j,i} x_i \right) \right)$$

for $k = 1, \dots, q$. One may choose with $\sigma_2(\alpha) = \alpha$, $\sigma_1(\alpha) = 1 / (1 + \exp(-\alpha))$. The only left here is to learn unknown parameters $\{w_{(\cdot),\cdot,\cdot}\}$.

2. FEEDFORWARD NEURAL NETWORK (MATHEMATICAL SET-UP)

Note 5. An (Artificial) Neural Network (NN) is an interconnection of several sigmoid or threshold functions associated arranged in layers, such that the outputs of one layer form the input of the next layer. Formally the structure of these interconnections can be depicted as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose nodes \mathcal{V} correspond to vertices/neurons and edges \mathcal{E} correspond to links between them.

Note 6. A Feed-Forward Neural Network (FFNN) or else multi-layer perceptron is a special case of NN whose vertices can be numbered so that all connections go from a neuron (vertex) to one with a higher number. Hence, neurons (vertices) have one-way connections to other neurons such that the output of a lower numbered neuron feeds the input of the higher numbered neuron (in a forward manner). The neurons can be arranged in layers so that connections go from one layer to a later layer. FFNN can be depicted by a directed acyclic graph², $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. (See Figure 2.1).

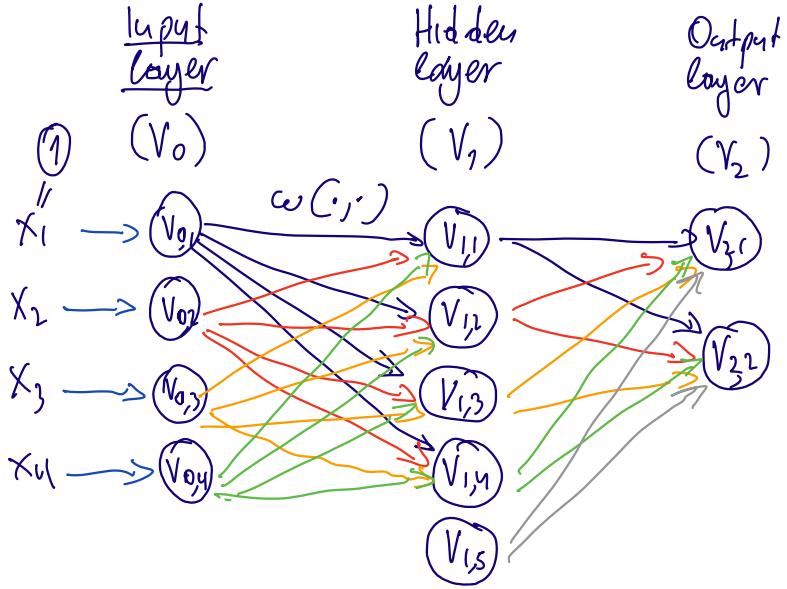


FIGURE 2.1. Feed forward neural network (1 hidden layer)

Note 7. We assume that the network is organized in layers. The set of nodes is decomposed into a union of (nonempty) disjoint subsets

$$V = \cup_{t=0}^T V_t$$

such that every edge in \mathcal{E} connects a node from V_t to a node from V_{t+1} , for $t = 1, \dots, T$.

Note 8. The first layer V_0 is called **input layer**. If x has d dimensions, then the first layer V_0 contains d nodes.

Note 9. The last layer V_T is called **output layer**. If y has q dimensions, then the last layer V_T contains q nodes.

Note 10. The intermediate layers $\{V_1, \dots, V_{T-1}\}$ are called **hidden layers**.

Note 11. In a neural network, the nodes of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ correspond to neurons.

Notation 12. The **i-th neuron of the t-th layer** is denoted as $v_{t,i}$.

Note 13. The output of neuron i in the input layer V_0 is simply x_i that is $o_{0,i}(x) = x_i$ for $i = \{1, \dots, d\}$.

Note 14. Each edge in the graph $(v_{t,j}, v_{t+1,i})$ links the output of some neuron $v_{t,j}$ to the input of another neuron $v_{t+1,i}$; i.e. $(v_{t,j}, v_{t+1,i}) \in \mathcal{E}$.

Note 15. We define a weight function $w : \mathcal{E} \rightarrow \mathbb{R}$ over the edges \mathcal{E} .

Note 16. Activation of neuron i at hidden layer 1 is the weighted sum of the outputs $o_{0,i}(x) = x_i$ of the neurons in V_0 which are connected to $v_{1,i}$ where weighting is according to function w , that is

$$(2.1) \quad \alpha_{1,i}(x) = \sum_{\forall j:(v_{0,j}, v_{1,i}) \in \mathcal{E}} w((v_{0,j}, v_{1,i})) x_j$$

Note 17. Each single neuron $v_{t,i}$ is modeled as a simple scalar function, $\sigma_t(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$, called **activation function** at layer t .

Note 18. We denote by $o_{t,i}(x) := \sigma_{t-1}(\alpha_{t-1,i}(x))$ **the output of neuron** $v_{t,i}$ when the network is fed with the input x .

Remark 19. Activation of neuron i at layer t is the weighted sum of the outputs $o_{t-1,j}(x)$ of the neurons in V_{t-1} which are connected to $v_{t,i}$ where weighting is according to function w , that is

$$(2.2) \quad \alpha_{t,i}(x) = \sum_{\forall j:(v_{t-1,j}, v_{t,i}) \in \mathcal{E}} w((v_{t-1,j}, v_{t,i})) o_{t-1,j}(x)$$

Note 20. The input of a neuron is obtained by taking a weighted sum of the outputs of all the neurons connected to it, where the weighting is according to w .

Note 21. The input to $v_{t+1,i}$ is activation $\alpha_{t+1,i}(x)$ namely a weighted sum of the outputs $o_{t,j}(x)$ of the neurons in V_t which are connected to $v_{t+1,i}$, where weighting is according to w . The output of $v_{t+1,i}$ is the application of the activation function $\sigma_{t+1}(\cdot)$ on its input $\alpha_{t+1,i}(x)$. –That's why it is called **Feed forward Neural Network**.

Summary 22. The feed-forward NN formula in a layer by layer manner is performed (defined) according to the following recursion.

At $t = 0$: for $i = 1, \dots, |V_0|$

$$o_{0,i}(x) := x_i$$

At $t = 0, \dots, T - 1$: for $i = 1, \dots, |V_{t+1}|$

$$\begin{aligned} \alpha_{t+1,i}(x) &= \sum_{\forall j:(v_{t,j}, v_{t+1,i}) \in \mathcal{E}} w((v_{t,j}, v_{t+1,i})) o_{t,j}(x) \\ o_{t+1,i}(x) &= \sigma_{t+1}(\alpha_{t+1,i}(x)) \end{aligned}$$

Note 23. Depth of the NN is the number of the layers **excluding the input layer**; i.e. T .

Note 24. Size of the network is the number $|V|$.

Note 25. Width of the NN is the number $\max_{\forall t}(|V_t|)$.

Note 26. The architecture of the neural network is defined by the triplet $(\mathcal{V}, \mathcal{E}, \sigma_t)$.

Note 27. The neural network can be fully specified by the quadruplet $(\mathcal{V}, \mathcal{E}, \sigma_t, w)$.

Example 28. Figure 2.1 denotes a NN with depth 2, size 11, width 5. The neuron with no incoming edges has $o_{1,5} = \sigma(0)$.

Notation 29. To ease the notation, we denote the weights as $w_{(t+1),i,j} := w((v_{t,j}, v_{t+1,i}))$. Using this notation, $w_{(t+1),i,j} = 0$ is equivalent in (2.2) to $(v_{t,j}, v_{t+1,i}) \notin \mathcal{E}$ and means that the link $v_{t,j} \rightarrow v_{t+1,i}$ is not in the network.

Note 30. Often a **constant neuron** $v_{t,0}$ (at each layer t and $i = 0$) which outputs 1; i.e. $o_{0,0}(x) = 1$ and $o_{t,0}(x) = 1$. The corresponding weight $w_{(t),k,0}$ is called **bias**. This resembles to the constant term in the linear regression.

Example 31. (Cont. Example 4) The 2 layer neural network

$$(2.3) \quad h_k(x) = \sigma_{(2)} \left(w_{(2),k,0} + \sum_{\forall j} w_{(2),k,j} \sigma_{(1)} \left(w_{(1),j,0} + \sum_{\forall i} w_{(1),j,i} x_i \right) \right)$$

can be written according to the recursion in Summary 22 as

- Input layer

$$o_{(0),i}(x) = \begin{cases} 1 & i = 0 \\ x_i & i = 1, \dots, d \end{cases}$$

- Hidden layer

$$\begin{aligned} \alpha_{(1),j}(x) &= w_{(1),j,0} + \sum_{\forall j} w_{(1),j,i} x_i \\ o_{(1),j}(x) &= \sigma_{(1)}(\alpha_{(1),j}(x)) \\ &= \sigma_{(1)} \left(w_{(1),j,0} + \sum_{\forall j} w_{(1),j,i} x_i \right) \end{aligned}$$

- Output layer

$$\begin{aligned} \alpha_{(2),k}(x) &= w_{(2),k,0} + \sum_{\forall j} w_{(2),k,j} o_{(1),j}(x) \\ o_{(2),k}(x) &= \sigma_{(2)}(\alpha_{(2),k}(x)) \\ &= \sigma_{(2)} \left(w_{(2),k,0} + \sum_{\forall j} w_{(2),k,j} o_{(1),j}(x) \right) \end{aligned}$$

If $h_k(x)$ returns values in \mathbb{R} , we can choose $\sigma_{(2)}$ as the identity function i.e., $\sigma_{(2)}(\alpha) = \alpha$. Note that (2.3) is also presented more compact

$$(2.4) \quad h_k(x) = \sigma_{(2)} \left(\sum_{\forall j} w_{(2),k,j} \sigma_{(1)} \left(\sum_{\forall i} w_{(1),j,i} x_i \right) \right)$$

by considering the constant neuron associated with first x which is set equal to 1, e.g $x_1 = 1$, similar to the linear regression models.

Note 32. Activation functions σ_t are non-increasing functions often sigmoids or threshold functions. Their choice is problem-dependent. some examples

- Identity function: $\sigma(\alpha) = \alpha$ cannot be used in hidden layers
- Threshold sigmoid: $\sigma(\alpha) = 1 (\alpha > 0)$
- Logistic sigmoid: $\sigma(\alpha) = \frac{1}{1+\exp(-\alpha)}$
- Rectified linear unit: $\text{RELU}(\alpha) = \max(\alpha, 0)$

Example 33. Examples on the choice of the activation function at the output layer T :

- In the univariate regression problem with prediction rule $h(\cdot) \in \mathbb{R}$, and examples $z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, we can choose $\sigma_T(\alpha) = \alpha$ to get $h(\alpha) = \sigma_T(\alpha) = \alpha$.
- In the binary logistic regression problem with prediction rule $h(\cdot) \in [0, 1]$ and examples $z_i = (x_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$, we can choose $\sigma_T(\alpha) = \frac{1}{1+\exp(-\alpha)}$ to get $h(\alpha) = \sigma_T(\alpha) = \frac{\exp(\alpha)}{1+\exp(\alpha)} = \frac{\exp(\alpha)}{1+\exp(\alpha)}$.

3. LEARNING NEURAL NETWORKS

Note 34. Assume we are interested in a prediction rule $h_{\mathcal{V}, \mathcal{E}, \sigma, w} : \mathbb{R}^{|V_0|} \rightarrow \mathbb{R}^{|V_T|}$ which is modeled as a feed-forward Neural Network with $(\mathcal{V}, \mathcal{E}, \sigma, w)$; that is

$$h_{\mathcal{V}, \mathcal{E}, \sigma, w}(x) = o_T(x)$$

where $o_T = (o_{T,1}, \dots, o_{T,|V_T|})^\top$ is according to the summary 22.

Note 35. Learning the architecture $(\mathcal{V}, \mathcal{E}, \sigma)$ of a neural network is a model selection task (similar to the variable selection in linear regression).

Note 36. We assume that the architecture $(\mathcal{V}, \mathcal{E}, \sigma)$ of the neural network $(\mathcal{V}, \mathcal{E}, \sigma, w)$ is fixed (given), and that there is interest in learning the weight function $w : \mathcal{E} \rightarrow \mathbb{R}$ or equivalently in vector form the vector of weights $\{w_{(t+1),i,j}\}$ where $w_{(t+1),i,j} := w((v_{t,j}, v_{t+1,i}))$.

Note 37. The class of hypotheses is

$$\mathcal{H}_{\mathcal{V}, \mathcal{E}, \sigma} = \{h_{\mathcal{V}, \mathcal{E}, \sigma, w} : \text{for all } w : \mathcal{E} \rightarrow \mathbb{R}\}$$

for given $(\mathcal{V}, \mathcal{E}, \sigma)$.

Notation 38. To simplify notation we will use h_w instead of $h_{\mathcal{V}, \mathcal{E}, \sigma, w}$.

Note 39. Assume that there is available a training set of examples (data-set) $\mathcal{S} = \{z_i = (x_i, y_i); i = 1, \dots, n\}$ with $x_i \in \mathcal{X} = \mathbb{R}^{|V_0|}$ and $y_i \in \mathcal{Y}$.

Note 40. To learn the unknown w , we need to specify a loss function $\ell(w, z)$ at some value of weight vector $w \in \mathbb{R}^{|\mathcal{E}|}$ and at some example $z = (x, y)$.

Example 41. For instance, loss function can be

- (1) $\ell(w, z) = \frac{1}{2} \|h_w(x) - y\|_2^2$ based on a norm
- (2) $\ell(w, z) = -\log(f(y|w))$ based on a pdf/pmf of the sampling distribution $f(y|w)$ of the target y given the weight vector w ; eg in a regression problem $y \sim N(h_w(x), \sigma^2)$.

Definition 42. Error function is a performance measure that can be defined as

$$(3.1) \quad \text{EF}(w | \{z_i^*\}) = \sum_{i=1}^n \ell(w, z_i^*)$$

where $\mathcal{S}^* = \{z_i^* = (x_i^*, y_i^*); i = 1, \dots, n^*\}$ is a set of examples which does not necessarily need to be the training set \mathcal{S} .

Example 43. Following we present some paradigms of NN implemented in applications seen before and with respect to quantities Notes 32, 34, 40, 42

Case 1. (Regression problem) Assume we wish to predict the mapping $x \xrightarrow{h(\cdot)} y$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Consider a predictive rule $h_w : \mathbb{R}^d \rightarrow \mathbb{R}$. Assume a training data-set $\{z_i = (x_i, y_i)\}$.

- The output activation function is the identity function $\sigma_T(a) = a$. This is because $h_w(x) = o_T(x) = \sigma_T(\alpha_T(x))$ by Summary 22 and Note 34. Since h_w returns in \mathbb{R} and the output activation $\alpha_T(x)$ returns in \mathbb{R} , then mapping $\sigma_T(a) = a$ suffices.
- A suitable loss can be

$$\ell(w, z = (x, y)) = \frac{1}{2} (h_w(x) - y)^2$$

Alternatively, if I consider a statistical model as

$$(3.2) \quad y_i | x, w \sim N(\mu_i, \beta^{-1}), \text{ where } \mu_i = h_w(x_i)$$

for some fixed $\beta > 0$, the loss can be

$$\begin{aligned} \ell(w, z = (x, y)) &= -\log(N(y | h_w(x), \beta^{-1})) \\ &= -\left(-\frac{1}{2} \log\left(\frac{1}{2\pi}\right) - \frac{1}{2} \log(\beta^{-1}) - \frac{1}{2} \frac{(h_w(x) - y)^2}{\beta^{-1}}\right) \\ &= \frac{1}{2} \beta (h_w(x) - y)^2 + \text{const...} \end{aligned}$$

- The Error function is

$$\text{EF}(w | z) = \sum_{i=1}^n \ell(w, z_i = (x_i, y_i)) = \frac{1}{2} \beta \sum_{i=1}^n (h_w(x_i) - y_i)^2$$

Case 2. (Multi-output regression problem) Assume we wish to predict the mapping $x \xrightarrow{h(\cdot)} y$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^q$. Consider a predictive rule $h_w : \mathbb{R}^d \rightarrow \mathbb{R}^q$. Assume a training data-set $\{z_i = (x_i, y_i)\}$.

- The output activation function is the identity function $\sigma_T(a) = a$. This is because $h_{w,k}(x) = o_{T,k}(x) = \sigma_T(\alpha_{T,k}(x))$ for $k = 1, \dots, q$ by Summary 22 and Note 34. Since h_w returns in \mathbb{R}^q and the output activation is $\alpha_T(x)$ in \mathbb{R} , then mapping $\sigma_T(a) = a$ suffices.

- A suitable loss can be

$$\ell(w, z = (x, y)) = \frac{1}{2} \|h_w(x) - y\|_2^2 = \frac{1}{2} \sum_{k=1}^q (h_{w,k}(x) - y_k)^2$$

Alternatively, if I consider a statistical model as

$$(3.3) \quad y_i|x_i, w \sim N(\mu, \Sigma), \text{ where } \mu_i = h_w(x_i)$$

for some fixed $\Sigma > 0$, the loss can be

$$\begin{aligned} \ell(w, z = (x, y)) &= -\log(N(y|h_w(x), \Sigma)) \\ &= -\left(-\frac{q}{2} \log\left(\frac{1}{2\pi}\right) - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (h_w(x) - y)^\top \Sigma^{-1} (h_w(x) - y)\right) \\ &= \frac{1}{2} (h_w(x) - y)^\top \Sigma^{-1} (h_w(x) - y) + \text{const...} \end{aligned}$$

- The Error function is

$$\begin{aligned} \text{EF}(w|z) &= \sum_{i=1}^n \ell(w, z_i = (x_i, y_i)) \\ &= \frac{1}{2} \sum_{i=1}^n (h_{w,k}(x_i) - y_{k,i})^\top \Sigma^{-1} (h_{w,k}(x_i) - y_{k,i}) \end{aligned}$$

where $y_{k,i}$ is the k -th dimension of the i -th example in the dataset.

Case 3. (Binary classification problem) Assume we wish to classify objects with features $x \in \mathbb{R}^d$ in 2 categories. Consider a predictive rule $h_w : \mathbb{R}^d \rightarrow (0, 1)$ as a classification probability i.e., $h_w(x) = \Pr(x \text{ belongs to class 1})$. Assume a training data-set $\{z_i = (x_i, y_i)\}$ with $y_i \in \{0, \dots, q\}$ labeling the class.

- A suitable output activation function can be the logistic sigmoid

$$\sigma_T(a) = \frac{1}{1 + \exp(-a)}$$

This is because $h_w(x) = o_T(x) = \sigma_T(\alpha_T(x))$ by Summary 22 and Note 34. Since h_w returns in $(0, 1)$ and the output activation is $\alpha_T(x)$ in \mathbb{R} , then the aforesaid mapping suffices.

- If I consider a statistical model as

$$(3.4) \quad y_i|x_i, w \sim \text{Bernoulli}(p_i), \text{ where } p_i = h_w(x_i)$$

with mass function

$$f(y|x, w) = h_w(x)^y (1 - h_w(x))^{1-y}$$

the loss can be

$$\begin{aligned} \ell(w, z = (x, y)) &= -\log(\text{Bernoulli}(y_i|h_w(x))) \\ &= -(y \log(h_w(x)) + (1 - y)(1 - \log(h_w(x)))) \end{aligned}$$

- The Error function given a set of examples $\{z_i^* = (x_i^*, y_i^*)\}$ is

$$\begin{aligned}\text{EF}(w|z^*) &= \sum_{i=1}^n \ell(w, z_i^* = (x_i^*, y_i^*)) \\ &= - \sum_{i=1}^n y_i^* \log(h_w(x_i^*)) - (1 - y_i^*)(1 - \log(h_w(x_i^*)))\end{aligned}$$

Case 4. (Multi-class classification problem) Assume we wish to classify objects with features $x \in \mathbb{R}^d$ in q categories. Consider a predictive rule $h_w : \mathbb{R}^d \rightarrow \mathcal{P}$, with $\mathcal{P} = \{\varpi \in (0, 1)^q : \sum_{j=1}^q \varpi_j = 1\}$ and $h_w = (h_{w,1}, \dots, h_{w,q})^\top$, as a classification probability i.e., $h_{w,k}(x) = \Pr(x \text{ belongs to class } k)$.

Assume a training data-set $\{z_i = (x_i, y_i)\}$ with $y_i \in \{0, 1\}$ labeling the class. Then it is

- A suitable output activation function can be the softmax function

$$(3.5) \quad \sigma_T(a_k) = \frac{\exp(a_k)}{\sum_{k'=1}^q \exp(a_{k'})}, \text{ for } k = 1, \dots, q$$

This is because $h_{w,k}(x) = o_{T,k}(x) = \sigma_T(\alpha_{T,k}(x))$ by Summary 22 and Note 34. Since h_w is essentially a probability vector and the output activation is $\alpha_{T,k}(x)$ in \mathbb{R} the aforesaid mapping suffices. Unfortunately, 3.5 is invariant to additive transformations $a \leftarrow a + c$ i.e., $\sigma_T(a_k) = \sigma_T(a_k + \text{const})$.

- Another suitable output activation function can be

$$(3.6) \quad \tilde{\sigma}_T(a_k) = \frac{\exp(a_k)}{1 + \sum_{k'=1}^{q-1} \exp(a_{k'})}, \text{ for } k = 1, \dots, q-1$$

This is because $h_{w,k}(x) = o_{T,k}(x) = \sigma_T(\alpha_{T,k}(x))$ by Summary 22 and Note 34. Since h_w is essentially a probability vector and the output activation is $\alpha_{T,k}(x)$ in \mathbb{R} the aforesaid mapping suffices as $h_{w,k}(x) = o_{T,k}(x)$ for $k = 1, \dots, q-1$ and $h_{w,q}(x) = 1 - \sum_{k=1}^{q-1} h_{w,k}(x)$. The advantage of (3.6) compared to (3.5) is that the former uses one less output neurons (less unknown parameters to learn). Also (3.6) is not invariant to additive transformations $a \leftarrow a + c$ unlike 3.5 i.e., $\tilde{\sigma}_T(a_k) \neq \tilde{\sigma}_T(a_k + \text{const})$.

- If I consider a statistical model as

$$y_i|x, w \sim \text{Multinomial}(p_i), \text{ where } p_i = h_w(x_i)$$

with mass function

$$f(y_i|x, w) = \prod_{k=1}^q h_{w,k}(x_i)^{y_{i,k}}$$

The loss can be

$$\begin{aligned}\ell(w, z = (x, y)) &= -\log(\text{Multinomial}(y|h_w(x))) \\ &= - \sum_{k=1}^q y_k \log(h_{w,k}(x))\end{aligned}$$

- The Error function given a set of examples $\{z_i^* = (x_i^*, y_i^*)\}$ is

$$\begin{aligned} \text{EF}(w|z^*) &= \sum_{i=1}^n \ell(w, z_i^* = (x_i^*, y_i^*)) \\ &= - \sum_{i=1}^n \sum_{k=1}^q y_{k,i} \log(h_{w,k}(x_i)) \end{aligned}$$

where $y_{k,i}$ is the k -th dimension of the i -th example in the dataset.

4. CLASSICAL LEARNING OF NEURAL NETWORK

Note 44. Our purpose is to find optimal $h_w \in \mathcal{H}_{\mathcal{V}, \mathcal{E}, \sigma}$ under loss $\ell(\cdot, \cdot)$ and against training data-set $\mathcal{S} = \{z_i = (x_i, y_i); i = 1, \dots, n\}$.

4.1. Standard. Essentially, this is an optimization problem, where the objective is to minimize either the risk function $R_g(w)$ or the empirical risk function $\hat{R}_S(w)$.

Problem 45. Compute optimal $w^* \in \mathbb{R}^{|\mathcal{E}|}$ by minimizing the risk function $R_g(w)$

$$(4.1) \quad w^* = \arg \min_{w \in \mathcal{H}} (R_g(w)) = \arg \min_{w \in \mathcal{H}} (\mathbb{E}_{z \sim g} (\ell(w, z)))$$

Problem 46. Compute optimal $w^* \in \mathbb{R}^{|\mathcal{E}|}$ by minimizing the empirical risk function $\hat{R}_S(w)$

$$(4.2) \quad w^* = \arg \min_{w \in \mathcal{H}} (\hat{R}_S(w)) = \arg \min_{w \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \ell(w, z_i) \right)$$

4.2. Regularization.

Note 47. Often neural network models are over-parameterized, in the sense that the dimensionality of $w \in \mathbb{R}^{|\mathcal{E}|}$ is too large, with negative consequences in its predictability. This can be addressed by learning the architecture NN, precisely by learning which of the edges of the FFNN significantly contribute to the NN model and should be kept, and which do not contribute and may be removed (or be inactive).

Note 48. To address Note 47, one can resort to shrinkage methods e.g., LASSO, Ridge, which indirectly allow edge/weight selection/elimination by shrinking the values of the weights $\{w_{(t),i,j}\}$ towards zero, and setting some of them as $w_{(t),i,j} = 0$ if their absolute value is small enough. This is based on the observation in (2.2) i.e., $w_{(t),i,j} = 0$ is equivalent to $(v_{t,j}, v_{t+1,i}) \notin \mathcal{E}$ which implies that the link $v_{t,j} \rightarrow v_{t+1,i}$ is not active (essentially not in the neural network).

Problem 49. Find $h_w \in \mathcal{H}_{\mathcal{V}, \mathcal{E}, \sigma}$ (essentially compute $w \in \mathbb{R}^{|\mathcal{E}|}$) under loss $\ell(\cdot, \cdot)$ and shrinkage term (or weight decay) $J(w; \lambda)$, and against training data-set $\mathcal{S} = \{z_i = (x_i, y_i); i = 1, \dots, n\}$. Compute $w^* \in \mathbb{R}^{|\mathcal{E}|}$, according to the minimization:

$$(4.3) \quad w^* = \arg \min_{w \in \mathcal{H}} (R_g(w) + J(w; \lambda))$$

$$(4.4) \quad = \arg \min_{w \in \mathcal{H}} (\mathbb{E}_{z \sim g} (\ell(w, z) + J(w; \lambda)))$$

and set $w_{t,i,j}^* = 0$ if $w_{t,i,j}^*$ is less than a threshold user specific value $\xi > 0$ i.e. $|w_{t,i,j}^*| < \xi$.

Note 50. Popular shrinkage terms $J(w; \lambda)$ are

- Ridge: $J(w; \lambda) = \lambda \|w\|_1$
- LASSO: $J(w; \lambda) = \lambda \|w\|_2^2$
- Elastic net: $J(w; \lambda = (\lambda_1, \lambda_2)) = \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$

Term 1

Term 1

5. STOCHASTIC GRADIENT IMPLEMENTATION FOR TRAINING NN

Note 51. Training a neural network model is usually a high-dimensional problem (essentially w has high dimensionality to provide a better approximation) and a big-data problem (essentially we need a large number of training examples to learn a large number of weights). For this reason, Stochastic Gradient Descent (and its variations) is a suitable stochastic learning algorithm.

Note 52. To address Problem 45, the recursion of the SGD with batch size m is

$$w^{(t+1)} = w^{(t)} - \eta_t \frac{1}{m} \sum_{j=1}^m \partial_w \ell \left(w^{(t)}, z_j^{(t)} \right)$$

Note 53. To address the Problem 49, the recursion of the SGD with batch size m is

$$w^{(t+1)} = w^{(t)} - \eta_t \left[\frac{1}{m} \sum_{j=1}^m \partial_w \ell \left(w^{(t)}, z_j^{(t)} \right) + \partial_w J(w; \lambda) \right]$$

for some positive λ which is user specified, or chosen via cross validation.

Note 54. The learning problem associated to the Neural network model is (almost always) non-convex due to the non-convex loss with respect to the w 's. Upon implementing SGD in the learning problem of Neural Network, the theoretical results in Section 4 (Handout 2) and Section 4 (Handout 3) will not be effective due to the violation of the assumptions.

Note 55. Practical guidelines for the use of SGD in the learning problem of NN:

- (1) Utilize a learning rate η_t that changes over the iterations. The choice of the sequence η_t is more significant. In practice, it is tuned by a trial and error manner: given a validation data-set \mathcal{S}^* you may perform cross validation based on Error Function (3.1).
- (2) Re-run the SGD procedure several times and by using different settings (learning rate η_t , batch size m) and different seed $w^{(0)}$ (randomly chosen) each time. Possibly, by luck, at some trial, we will initialize the SGD process with a random seed $w^{(0)}$ producing a trace leading to a good local minimum w^* .
- (3) The output w_{SGD}^* returned by SGD is the best discovered w tested by using a performance measure (Error Function) using a validation set $\mathcal{S}^* = \{z_i^* = (x_i^*, y_i^*); i = 1, \dots, n^*\}$; Eg.

$$w_{\text{SGD}}^* = \arg \min_{w^{(t)}} \left(\text{EF} \left(w^{(t)} | \{z_i^*\} \right) \right).$$

5.1. Error backpropagation (to compute $\nabla_w \ell(w, z)$).

Note 56. The error backpropagation procedure is an efficient algorithm for the computation of the gradient $\nabla_w \ell(w, z)$ of the loss function $\ell(w, z)$ at some value of w and some example $z = (x, y)$ as required for the implementation of stochastic gradient based algorithm for the learning problems under consideration.

Assumption 57. *Error backpropagation assumes that $\ell(w, z)$ is differentiable at w for each value of z ; $\nabla_w \ell(w, z)$ exists.*

Notation 58. Let $V_t = \{v_{t,1}, \dots, v_{t,k_t}\}$ be the t -th layer of a NN and $k_t = |V_t|$ the number of neuron in layer $t = 1, \dots, T$.

Notation 59. Let $o_t = (o_{t,1}, \dots, o_{t,k_t})^\top$ be the vector of the outputs of the t -th layer of a NN.

Notation 60. Let $\alpha_t = (\alpha_{t,1}, \dots, \alpha_{t,k_t})^\top$ be the vector of the activations of the t -th layer of a NN.

Notation 61. We denote as $\ell_t(\cdot)$ the loss function $\ell(w, z)$ as a function of the output o_t at t -th layer.

- E.g it is $\ell_T(\xi) = \frac{1}{2}(\xi - y)^2$ if $\ell(w, z) = \frac{1}{2}(h_w(x) - y)^2$ since $h_w(x) = o_T(x)$.

Algorithm 62. (*Error backpropagation*)

Requires: The NN $(\mathcal{V}, \mathcal{E}, \sigma, w)$ with the values of the a weight vector $w \in \mathbb{R}^{|\mathcal{E}|}$, and example value $z = (x, y)$

Returns: $\nabla_w \ell(w, z) = \left(\frac{\partial}{\partial w_{t,i,j}} \ell(w, z); \forall t, i, j \right)$

Initialize:

$$\text{Set } w_{t+1,j,i} = \begin{cases} w(v_{t,j}, v_{t+1,j}) & \text{if } (v_{t,j}, v_{t+1,j}) \in \mathcal{E} \\ 0 & \text{if } (v_{t,j}, v_{t+1,j}) \notin \mathcal{E} \end{cases}$$

Forward pass:

(1) For $i = 1, \dots, d$

Set:

$$o_{0,i} = x_i$$

(2) For $t = 1, \dots, T$

For $i = 1, \dots, k_t$

Compute:

$$\alpha_{t,i} = \sum_{j=1}^{k_{t-1}} w_{t,i,j} o_{t-1,j}$$

Compute:

$$o_{t,i} = \sigma_t(\alpha_{t,i})$$

Backward pass:

(1) Set^a:

$$\delta_T = \frac{d\ell_T}{do_T}(o_T)$$

(2) For $t = T - 1, \dots, 1$

For $i = 1, \dots, k_t$

Compute:

$$\delta_{t,i} = \sum_{j=1}^{k_{t+1}} w_{t+1,j,i} \delta_{t+1,j} \left. \frac{d}{da} \sigma_{t+1}(a) \right|_{a=\alpha_{t+1,j}}$$

Output:

for each edge $(v_{t-1,j}, v_{t,i}) \in \mathcal{E}$

Set,

$$\frac{\partial}{\partial w_{t,i,j}} \ell(w, z) = \delta_{t,i} \sigma'_t(\alpha_{t,i}) o_{t-1,j}$$

for $i = 1, \dots, k_t$ and $j = 1, \dots, k_{t-1}$.

^aEg if $\ell(w, z) = \frac{1}{2} \sum_{k=1}^q (h_k(x) - y_k)^2$ then $\delta_T = \frac{d\ell_T}{do_T}(o_T) = o_T - y$, or otherwise $\frac{d\ell_T}{do_{T,k}}(o_T) = o_{T,k} - y_k$ for all $k = 1, \dots, q$

Remark 63. Several sources in the literature use the notation

$$\tilde{\delta}_t := \frac{d\ell_t}{d\alpha_t} = \frac{d\ell_t}{do_t} \frac{do_t}{d\alpha_t}$$

instead of

$$\delta_t = \frac{d\ell_t}{do_t}$$

used in Algorithm 62. Then the backward pass and output steps in Algorithm 62 can be equivalently restated as

Backward pass:

(1) For $t = T$,

Set:

$$\tilde{\delta}_T = \frac{d\ell_T}{d\alpha_T} (o_T)$$

(2) For $t = T - 1, \dots, 1$

For $i = 1, \dots, k_t$

Compute:

$$\tilde{\delta}_{t,i} = \sigma'_t(\alpha_{t,k}) \sum_{j=1}^{k_{t+1}} w_{t+1,j,i} \tilde{\delta}_{t+1,j}$$

Output:

- for each edge $(v_{t-1,j}, v_{t,i}) \in \mathcal{E}$

Set

$$\frac{\partial}{\partial w_{t,i,j}} \ell(w, z) = \tilde{\delta}_{t,i} o_{t-1,j}$$

Example 64. Consider the multi-output regression problem, assume a predictive rule $h_w : \mathbb{R}^d \rightarrow \mathbb{R}^q$ with $h_w = (h_{w,1}, \dots, h_{w,q})$ and

$$h_k(x) = \sigma_2 \left(\sum_{j=1}^c w_{2,k,j} \sigma_1 \left(\sum_{i=1}^d w_{1,j,i} x_i \right) \right)$$

with activation functions $\sigma_2(a) = a$, and $\sigma_1(a) = \tanh(a) = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}$, and loss $\ell(w, z) = \frac{1}{2} \sum_{k=1}^q (h_k(x) - y_k)^2$ for example $z = (x, y)$. Perform the error back propagation steps to compute the elements of $\nabla_w \ell(w, z)$ at some w and z .

Solution.

Forward pass:

Set: $o_{0,i} = x_i$ for $i = 1, \dots, d$

Compute:

at $t = 1$: for $j = 1, \dots, c$

comp: $\alpha_{1,j} = \sum_{i=1}^d w_{1,j,i} x_i$

comp: $o_{1,j} = \tanh(\alpha_{1,j})$

at $t = 2$: for $k = 1, \dots, q$

comp: $\alpha_{2,k} = \sum_{j=1}^c w_{2,k,j} o_{1,j}$

comp: $o_{2,k} = \alpha_{2,k}$

get: $h_k = o_{2,k}$

Note that $\frac{d}{d\xi}\sigma_1(\xi) = 1 - (\sigma_1(\xi))^2$ and that $\frac{d}{d\xi}\sigma_2(\xi) = 1$.

Backward pass:

at $t = 2$: for $k = 1, \dots, q$

comp:

$$\tilde{\delta}_{2,k} = \frac{\partial}{\partial \alpha_{2,k}} \ell_T = \sum_{j=1}^q \frac{\partial \ell_T}{\partial o_{2,j}} (o_{2,j}) \frac{\partial o_{2,j}}{\partial \alpha_{2,k}} (\alpha_{2,k}) = \frac{\partial \ell_T}{\partial o_{2,k}} (o_{2,k}) \frac{\partial o_{2,k}}{\partial \alpha_{2,k}} (\alpha_{2,k}) = h_k - y_k$$

at $t = 1$: for $j = 1, \dots, c$

comp:

$$\begin{aligned} \tilde{\delta}_{1,j} &= \left. \frac{d}{d\xi} \sigma_1(\xi) \right|_{\xi=\alpha_{1,j}} \sum_{k=1}^q w_{2,j,k} \tilde{\delta}_{2,k} \\ &= \left(1 - (o_{1,j})^2 \right) \sum_{k=1}^q w_{2,j,k} \tilde{\delta}_{2,k} \end{aligned}$$

Output:

$$\frac{\partial \ell(w, z)}{\partial w_{1,j,i}} = \tilde{\delta}_{1,j} x_i \text{ and } \frac{\partial \ell(w, z)}{\partial w_{2,k,j}} = \tilde{\delta}_{2,k} o_{2,j}$$

5.2. Design of Error back-propagation (Algorithm 62).

Notation 65. Let $\ell_t : \mathbb{R}^{k_t} \rightarrow \mathbb{R}$ be the loss function of the sub-network defined by layers $\{V_1, \dots, V_T\}$ as a function of the outputs o_t of the neurons in V_t .

Notation 66. Let W_t be the $k_{t-1} \times k_t$ matrix of the weights of the t -th layer of a NN such as $[W_t]_{i,j} := w_{t,i,j} = w(v_{t-1,j}, v_{t,i})$.

Notation 67. We introduce notation, such that if $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ and $a \in \mathbb{R}^d$, then $\sigma(a) \in \mathbb{R}^d$ is a d -dimensional vector such that $\sigma(a) = (\sigma(a_1), \dots, \sigma(a_d))$.

Notation 68. The vector of the outputs $o_t \in \mathbb{R}^{k_t}$ of the neurons of V_t can be written as $o_t = \sigma_t(\alpha_t)$; aka $o_{t,j} = \sigma_t(\alpha_{t,j})$, for $j = 1, \dots, k_t$.

Notation 69. The vector of activations, aka vector of the inputs, of the neurons of V_t can be written as $\alpha_t = W_t o_{t-1}$.

Note 70. Hence,

$$(5.1) \quad \ell_t(o_t) = \ell_t(\sigma_t(\alpha_t)) = \ell_t(\sigma_t(W_t o_{t-1}))$$

Notation 71. To facilitate differential calculus, let O_{t-1} be the $k_t \times (k_{t-1} k_t)$ matrix

$$O_{t-1} = \begin{bmatrix} o_{t-1}^\top & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & o_{t-1}^\top \end{bmatrix}$$

and let $w_t \in \mathbb{R}^{k_{t-1}k_t}$ is the vector $w_t = (W_{t,1}, \dots, W_{t,k_{t-1}})^\top$ by concatenating the rows of matrix W_t .
Hence

$$(5.2) \quad W_t o_{t-1} = O_{t-1} w_{t-1}$$

Note 72. Hence, from (5.1) and (5.2), it is

$$\ell_t(o_t) = \ell_t \left(\sigma_t \left(\underbrace{O_{t-1} w_t}_{=o_t} \right) \right)$$

Note 73. By chain rule it is

$$(5.3) \quad \begin{aligned} \frac{d}{dw_t} \ell_t(o_t) &= \frac{d\ell_t}{do_t} \frac{do_t}{d\alpha_t} \frac{d\alpha_t}{dw_t} \\ &= \frac{d\ell_t}{do_t} \text{diag} \left(\frac{d}{d\xi} \sigma_t(\xi) \Big|_{\xi=\alpha_t} \right) O_{t-1} \end{aligned}$$

Notation 74. Let as define

$$\delta_t = \frac{d}{do_t} \ell_t(o_t)$$

and name them as ‘errors’.

Note 75. Then (5.3), becomes

$$\begin{aligned} \frac{d\ell_t}{dw_t} &= \delta_t \text{diag} \left(\frac{d}{d\xi} \sigma_t(\xi) \Big|_{\xi=\alpha_t} \right) O_{t-1} \\ &= \left(\delta_{t,1} \sigma'(\alpha_{t,1}) o_{t-1,1}^\top, \dots, \delta_{t,k_t} \sigma'(\alpha_{t,k_t}) o_{t-1,k_t}^\top \right) \end{aligned}$$

Note 76. I will find a way to compute $\{\delta_t\}$ recursively from V_T to V_0 . For $t = T$, it is

$$\delta_T = \frac{d}{do_T} \ell_T(o_T) = \frac{d}{do_T} \ell(o_T, z)$$

Note that

$$\ell_t(o_t) = \ell_{t+1} \left(\sigma_{t+1} \left(\underbrace{W_{t+1} o_t}_{=o_{t+1}} \right) \right)$$

Then for $t = T-1, \dots, 1$, it is

$$\begin{aligned} \delta_t &= \frac{d\ell_t}{do_t} = \frac{d\ell_{t+1}}{do_t} \\ &= \frac{d\ell_{t+1}}{do_{t+1}} \frac{do_{t+1}}{d\alpha_{t+1}} \frac{d\alpha_{t+1}}{do_t} \\ &= \delta_{t+1} \text{diag} \left(\frac{d}{d\xi} \sigma_{t+1}(\xi) \Big|_{\xi=\alpha_{t+1}} \right) W_{t+1} \end{aligned}$$

Note 77. Hence, for $t = 1, \dots, T$ it is

$$\frac{d\ell(w, z)}{dw_t} = \delta_t \text{diag} \left(\left. \frac{d}{d\xi} \sigma_t(\xi) \right|_{\xi=\alpha_t} \right) O_{t-1}$$

or element wise, for $t = 1, \dots, T, j = 1, \dots, k_{t-1}, i = 1, \dots, k_t$ it is

$$\frac{\partial \ell(w, z)}{\partial w_{t,i,j}} = \delta_{t,i} \sigma'_t(\alpha_{t,i}) o_{t-1,j}$$

Remark 78. Error back-propagation idea can also be used to efficiently compute the gradient $\nabla_x \ell(w, z = (x, y))$ of the loss function $\ell(w, z = (x, y))$ with respect to the inputs x . The idea is the same, use chain rule to find a recursive procedure.

5.3. Preconditioning and computation of the Hessian.

Note 79. Recall (Handout 3, Algorithm 43), that SGD may be improved by using a suitable preconditioner $P_t > 0$ as

$$w^{(t+1)} = w^{(t)} - \eta_t P_t \nabla_w \ell(w^{(t)}, z^{(t)})$$

such a preconditioner can be the $P_t := [H_t + \epsilon I_d]^{-1}$ where H_t is the Hessian of $\ell(w^{(t)}, z^{(t)})$ and $\epsilon > 0$.

Note 80. The computation of the Hessian H_t can be done by using error propagation ideas for

$$[H_t]_{i,j} = \left. \frac{\partial^2}{\partial w_i \partial w_j} f(w) \right|_{w=w^{(t)}}$$

We will not go further to exact computations.

Note 81. The exact computation of the Hessian H_t of the loss $\ell(\cdot, z)$ in NN setting can be computationally expensive and hence approximations are often used (with hope to work well). One can implement the general purpose AdaGrad (Section 7.1, Handout 3). In what follows we present other alternatives tailored to the NN model.

Note 82. Consider the regression problem with predictive rule $h : \mathbb{R}^d \rightarrow \mathbb{R}$ with $h_w(x) = o_T(x) = \alpha_T(x)$, and loss function $\ell(w, z = (x, y)) = \frac{1}{2} (h_w(x) - y)^2$. Then the Hessian of ℓ is

$$(5.4) \quad \begin{aligned} H &= \frac{d(\nabla_w \ell)}{dw} = \frac{d}{dw} \left(\nabla_w h_w(x) (h_w(x) - y)^\top \right) \\ &= \left(\frac{d}{dw} \nabla_w h_w(x) \right) (h_w(x) - y)^\top + \nabla_w h_w(x) (\nabla_w h_w(x))^\top \end{aligned}$$

We can expect that $h_w(x) \approx y$ provided that the network is well trained due to the global approximation ability of the FFNN. Yet, we can expect that $h_w = E(y)$ provided that we train the network under the quadratic loss and because $\arg \min_h E_{y \sim g}(h - y)^2 = E_{y \sim g}(y)$. Hence a reasonable approximation can be

$$\begin{aligned} H &\approx \nabla_w h_w(x) (\nabla_w h_w(x))^\top \\ &\stackrel{\sigma_T(a)=a}{=} \nabla_w \alpha_T(x) (\nabla_w \alpha_T(x))^\top \end{aligned}$$

Consequently, the approximation of the Hessian of the Error function $\text{EF}(w | \{z_i\}) = \sum_{i=1}^n \ell(w, z_i = (x_i, y_i))$ is

$$(5.5) \quad H_n = \sum_{i=1}^n \frac{d(\nabla_w \text{EF})}{dw} \approx \sum_{i=1}^n \nabla_w h_w(x_i) (\nabla_w h_w(x_i))^{\top}$$

Note 83. (Cont. Note 82) To efficiently compute the inverse H_n^{-1} of (5.5) I can utilize Woodbury identity

$$\left(M + vv^{\top} \right)^{-1} = M^{-1} - \frac{(M^{-1}v)(v^{\top}M^{-1})}{1 + v^{\top}M^{-1}v}$$

offering a way to avoid the computational expensive task of directly inverting the high dimensional H_n . Given $v_i = \nabla_w h_w(x)$, it is

$$(5.6) \quad \begin{aligned} (H_n)^{-1} &= \left(\sum_{i=1}^n v_i v_i^{\top} \right)^{-1} = \left(\sum_{i=1}^{n-1} v_i v_i^{\top} + v_n v_n^{\top} \right)^{-1} = \left(H_{n-1} + v_n v_n^{\top} \right)^{-1} \\ &= H_{n-1}^{-1} - \frac{(H_{n-1}^{-1} v_n)(v_n^{\top} H_{n-1}^{-1})}{1 + v_n^{\top} H_{n-1}^{-1} v_n} \end{aligned}$$

In practice I start with $H_0 = \epsilon I$ with $\epsilon > 0$ small, and iterate (5.6).

Note 84. Likewise and by modifying (5.4), one can compute the corresponding approximations for the classification problems or problems with different loss functions.

6. BAYESIAN ARTIFICIAL NEURAL NETWORKS

Note 85. Consider a feed-forward Neural Network with $(\mathcal{V}, \mathcal{E}, \sigma, w)$. We use the same notation and structure as in the classical treatment above. Assume the architecture $(\mathcal{V}, \mathcal{E}, \sigma)$ of a neural network is fixed/known, and interest lies in learning the weights $\{w_{t,i,j}\}$.

Note 86. Assume there is available a training set of examples (data-set) $\mathcal{S} = \{z_i = (x_i, y_i); i = 1, \dots, n\}$ with $x_i \in \mathcal{X} = \mathbb{R}^{|V_0|}$ and $y_i \in \mathcal{Y}$. The Bayesian NN model is then

$$\begin{cases} y_i | x_i, w \stackrel{\text{ind}}{\sim} f(y_i | x_i, w), i = 1, \dots, n & \text{(sampling distribution)} \\ w \sim f(w) & \text{(prior distribution)} \end{cases}$$

where the sampling distribution is specified either according to the experimental design (how $\{z_i\}$ are collected or based on subjective judgments, while the prior distribution is specified based on subjective manner.

Note 87. In most of the real applications, it is difficult (almost impossible) to specify the sampling distribution based on the experimental design or judgments, and almost impossible to specify the prior of the weights based on subjective judgments.

Note 88. One could specify the sampling distribution based on the loss function $\ell(h_w(x), z = (x, y))$ as

$$(6.1) \quad f(y_i | x_i, w) \propto \exp(-\ell(h_w(x_i), (x_i, y_i))),$$

No need to
memorize
Woodbury
identity

based on the argument that sampling distribution in the posterior distribution contraindicates how far the theoretical model $h_w(x_i)$ (as casted in a NN) is from the corresponding observation y_i via the likelihood.

Note 89. The prior density of the weights may be specified based on some shrinkage term $J(w; \lambda)$ (Note 50) as

$$(6.2) \quad f(w) \propto \exp(-J(w; \lambda))$$

based on the argument that I often model the predictive rule $h_w(\cdot)$ with an over-parametrized NN (with many weights, more than needed) and many of them should be shrunk to zero. Also as we see in Note 97, careless training of NN tends to produce over-fitted NN with large weights (in abs values).

Note 90. Regarding prior it is often, $w \sim N(0, I\lambda^{-1})$ for some small λ controlling prior uncertainty about weights, i.e. 50% chances for the weight to be above or below zero.

Note 91. The corresponding posterior (according to the Bayes theorem) is

$$f(w| \{z_i\}) = \frac{\prod_{i=1}^n f(z_i|w) f(w)}{\int \prod_{i=1}^n f(z_i|w') f(w') dw'} \propto \prod_{i=1}^n f(z_i|w) f(w)$$

and given 6.1 and 6.2, in log scale I get

$$\log(f(w| \{z_i\})) = - \sum_{i=1}^n \ell(h_w(x_i), (x_i, y_i)) - J(w; \lambda) + \text{const}$$

that resembles the EF with a shrinkage term in classical learning.

Example 92. (Cont. Example 43)

Case 1. (Regression problem) Sampling distribution can be specified as in (6.1) based on loss

$$\ell(w, z = (x, y)) = \frac{\beta}{2} (h_w(x) - y)^2$$

Note this is equivalent to as if we had considered a sampling distribution

$$(6.3) \quad y_i|x_i, w \sim N(\mu_i, \beta^{-1}), \text{ where } \mu_i = h_w(x_i)$$

for some fixed $\beta > 0$ based on (3.2). Prior can be specified with density (6.2) where

$$J(w; \lambda) = \frac{\lambda}{2} \|w\|_2^2$$

is the Ridge shrinkage term. Note this is equivalent to as if we had considered

$$(6.4) \quad w \sim N(0, I\lambda^{-1})$$

The resulted posterior density is

$$(6.5) \quad f(w| \{z_i\}) \propto \exp\left(-\frac{\beta}{2} \sum_{i=1}^n (h_w(x_i) - y_i)^2 - \frac{\lambda}{2} \|w\|_2^2\right)$$

Case 2. (Multi-output regression problem) Try to do it by yourself

Case 3. (Binary classification problem) Sampling distribution can be specified as in (6.1) based on loss

$$\ell(w, (x, y)) = y \log(h_w(x)) + (1 - y)(1 - \log(h_w(x)))$$

which is equivalent to as if we had considered a sampling distribution

$$y_i | x_i, w \sim \text{Bernoulli}(p_i), \text{ where } p_i = h_w(x_i)$$

based on (3.4). Prior can be specified as in (6.2) with the LASSO shrinkage term

$$J(w; \lambda) = \lambda \|w\|_1;$$

The resulted posterior density is

$$f(w | \{z_i\}) \propto \exp \left(- \sum_{i=1}^n [y_i \log(h_w(x_i)) + (1 - y_i)(1 - \log(h_w(x_i)))] - \lambda \|w\|_1 \right)$$

Case 4. (Multi-class classification problem) Try to do it by yourself.

Note 93. Sampling from the posterior can be performed via SGLD due to the high-dimensionality in the weights w and the big size of the training data set. Recall the recursion of the SGLD with batch size m is

$$(6.6) \quad w^{(t+1)} = w^{(t)} + \eta_t \left(\frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log \left(f(z_j^{(t)} | w^{(t)}) \right) + \nabla_w \log \left(f(w^{(t)}) \right) \right) + \sqrt{\eta_t} \sqrt{\tau} \epsilon_t,$$

for $\epsilon_t \sim N(0, 1)$. (See Algorithm 21 in Handout 4: Bayesian Learning via Stochastic gradient and Stochastic gradient Langevin dynamics).

Note 94. Given sample values $\{w^{(t)}\}_{t=1}^T$ produced from SGLD recursions (6.6), learning of any function $h_w(x)$ of the w 's can be performed via

- (1) Monte Carlo estimation namely averaging the samples values $\{w^{(t)}\}_{t=1}^T$ as

$$\widehat{h}_w(x) = \frac{1}{T} \sum_{t=1}^T h_{w^{(t)}}(x), \quad (\text{Monte Carlo estimator})$$

- (2) Maximum-A-posteriori (MAP) estimation namely find the best \hat{w}^* among the samples values $\{w^{(t)}\}_{t=1}^T$ that maximizes the posterior i.e

$$\hat{w}^* = \arg \max_{\{w^{(t)}\}_{t=1}^T} (\log f(\{z_i\} | w) + \log f(w))$$

and compute

$$\widehat{h}_w(x) = h_{\hat{w}^*}(x), \quad (\text{MAP estimator})$$

7. THEORETICAL ASPECTS

Note 95. We will not go this direction here. For the interested student, I suggest for starters, Ch 5 from “Ripley, B. D. (2007). Pattern recognition and neural networks. Cambridge university press.”;

and for advanced Ch 30 from “Devroye, L., Györfi, L., & Lugosi, G. (2013). A probabilistic theory of pattern recognition (Vol. 31). Springer Science & Business Media.”

8. COMMENTS, GUIDELINES, AND DISCUSSIONS

8.1. Over-fitting issues.

Note 96. Neural Networks can be “over-parametrized”, for instance by consisting of a large number of layers each of them having a large number of neurons able to represent each feature/characteristic of the pattern of interest to be learned. Careless training may produce NN models with bad generalization predictive properties.

Note 97. Training of non-linear (non-convex) NN models corresponds to an iterative reduction of the Error Function defined on the training data-set. It has been observed that the Error Function defined on the validation data-set (independent to the training data set) often shows a decrease at first, followed by an increase as the NN starts to overfit. This overfit is often associated to the production of weight values larger than needed to be generalised. It is desirable to avoid this over-fitting with purpose to obtain a NN with good generalization performance.

Note 98. **Early stopping** is a way of limiting the effective network complexity by halting training before a minimum of the training error has been reached. During training, we monitor the NN performance against the EF defined on the validation data-set, and stop the training procedure when just before the EF defined on the validation data set start increasing (aka before overfitting signs) to prevent the model memorizing too much information about the training set.

Note 99. **Regularization** as in Section 4.2 can alternatively be used to address the above issue by using a shrinkage term preventing the weights to grow too much away from zero.

8.2. Non-identifiabilities.

Note 100. The weights $\{w_{t,i,j}\}$ have non-identifiabilities. Consider the FFNN in Figure 8.2 with 1 hidden layer of M neurons, activation function $\sigma_1(a) = \tanh(a)$ and full connectivity in both layers. If we flip the sign of all of the w 's feeding into a particular hidden unit (for a given input pattern) then the sign of the activation of the hidden unit will be reversed because $\tanh(-a) = -\tanh(a)$. This transformation can be compensated by changing the sign of all of the w 's leading out of that hidden unit. By changing the signs of a particular group of w 's, the input-output mapping function represented by the FFNN is unchanged, and so there are two different weight vectors resulting the same mapping function. I can do M such flips (there are M hidden neurons) leading to 2^M equivalent parameter settings. Similarly, we can permute the labels of the neurons in hidden layer without changing the loss function; there are $M!$ such permutations. Hence the equivalent parameter settings in total are $2^M M!$. –This is not harmful in training or predictive rule computation as we just need to find one such parameter setting.

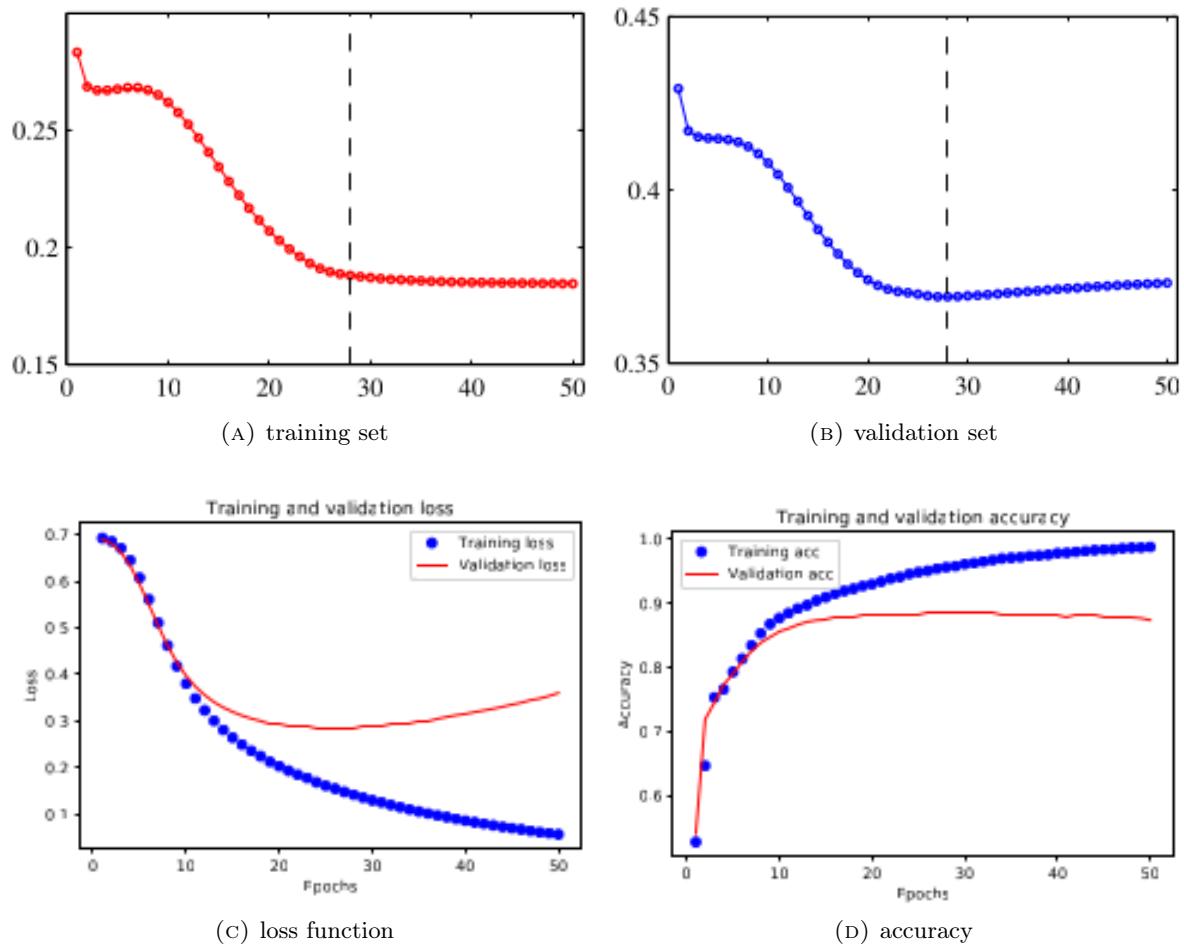


FIGURE 8.1. Behavior of the Error Function wrt the iterations, against a training set and against a validation set.

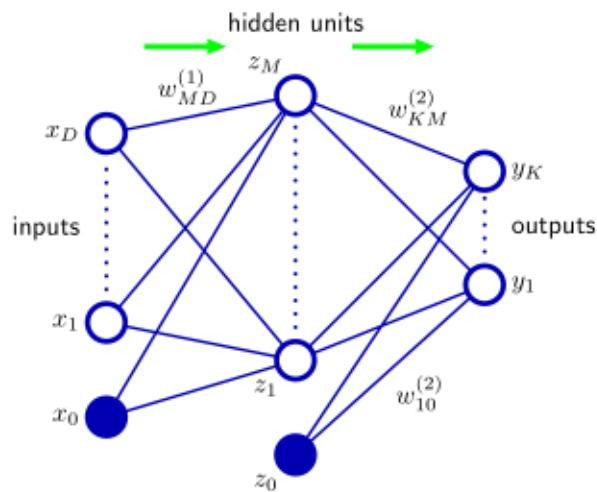


FIGURE 8.2. A FFNN

8.3. Non-convexity.

Note 101. Learning problems with NN are non-convex, hence the loss function to be minimized has many local minima. The consequence is that the SGD/SGLD learning algorithms may be trapped in a local optima and never reach any of the global ones. The number of local minima is exaggerated due to the symmetries discussed in Note 100. Due to the large number of data (big-data) used to train the NN (in real life) such local minima become more shallow while the global minima more picky. Due to this and the stochastic nature of SGD/SGLD algorithms, SGD/SGLD may be able (by chance) to escape from such local minima/maxima and reach the global ones.

Note 102. To address local optima issue, run SGD/SGLD learning algorithms multiple times by initializing them with different seeds each time.

8.4. Skip-layers 2.4.

Note 103. The general definition of feed forward Neural Network allows “skip-layer” connections that is the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in NN architecture includes edges which may not necessarily connect neurons between consecutive layers, namely $(v_{t,j}, v_{t+k,i}) \in \mathcal{E}$ for some $k > 1$.

Example 104. For instance, a fully connected FFNN with 1 hidden layer is

$$(8.1) \quad h_k(x) = \sigma_{(2)} \left(\underbrace{\sum_{\forall i} w_{(1),k,i}^{\text{skip}} x_i}_{\text{skip layer terms}} + \sum_{\forall j} w_{(2),k,j} \sigma_{(1)} \left(\sum_{\forall i} w_{(1),j,i} x_i \right) \right)$$

Compared to (2.4) in Example 31 and (2.4), (8.1) links the input to the output.

Note 105. In skip-layer FFNN cases, the error back propagation (Section 5.1) has to be adjusted properly.

Note 106. FFNN are mainly used without skip-layers because theory (Section 7) states that FFNN are global approximators even without skip layers, as well as because of computational inconvenience and modeling parsimony.

APPENDIX A. ABOUT GRAPHS

Definition 107. A directed graph is an ordered pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ comprising

- a set of nodes \mathcal{V} (or vertices), where nodes are abstract objects, and
- a set of edges $\mathcal{E} = \{(v, u) | v \in \mathcal{E}, u \in \mathcal{E}, v \neq u\}$ (or directed edges, directed links, arrows) which are ordered pairs of vertices (that is, an edge is associated with two distinct vertices).

Definition 108. An edge-weighted graph or a network is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ equipped with a weight function $w : \mathcal{E} \rightarrow \mathbb{R}$ that assigns a number (the weight) to each edge $e \in \mathcal{E}$.

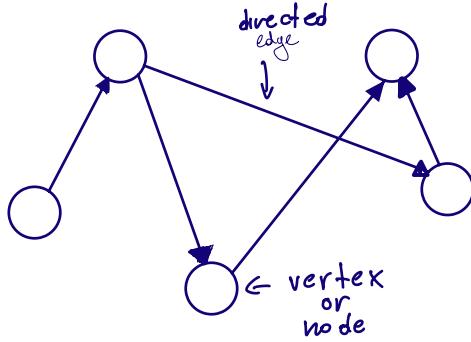


FIGURE A.1. A directed graph

APPENDIX B. ABOUT PARTIAL DERIVATIVES

Note 109. Consider a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

Definition 110. The partial derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at the point $a = (a_1, \dots, a_n) \in U \subseteq \mathbb{R}^n$ with respect to the i -th variable is denoted as

$$\frac{\partial f}{\partial x_i}(a) \text{ or } \left. \frac{\partial f}{\partial x_i}(x) \right|_{x=a}$$

and defined as

$$\begin{aligned} \left. \frac{\partial f}{\partial x_i}(x) \right|_{x=a} &= \lim_{h \rightarrow 0} \frac{f(a_1, \dots, a_{i-1}, a_i + h, a_{i+1}, \dots, a_n) - f(a_1, \dots, a_{i-1}, a_i, a_{i+1}, \dots, a_n)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(a + he_i) - f(a)}{h} \end{aligned}$$

where e_i is a $0 - 1$ vector with only one ace in the i -th location.

Remark 111. Essentially, Definition 110 says that the partial derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at the point $a = (a_1, \dots, a_n) \in U \subseteq \mathbb{R}^n$ with respect to the i -th variable is

$$\frac{\partial f}{\partial x_i}(a) = \left. \frac{dg}{dh}(h) \right|_{h=0}$$

the derivative of function $g(h) := f(a_1, \dots, a_{i-1}, a_i + h, a_{i+1}, \dots, a_n)$ at value 0.

Example 112. Consider a function f with $f(x_1, x_2) = x_1^2 + x_1 x_2^3$. Compute its partial derivatives at $a = (2, 3)^\top$; i.e. $\frac{\partial f}{\partial x_1}(a)$ and $\frac{\partial f}{\partial x_2}(a)$.

Solution. It is

$$\begin{aligned}\frac{\partial f}{\partial x_1}(a) &= \frac{\partial f}{\partial x_1}(x)\Big|_{x=a} = \frac{d}{dx_1}(x_1^2 + x_1 x_2^3)\Big|_{x=a} = 2x_1 + x_2^3|_{x=a} \\ &= 2a_1 + a_2^3 = 4 + 27 = 31 \\ \frac{\partial f}{\partial x_2}(a) &= \frac{\partial f}{\partial x_2}(x)\Big|_{x=a} = \frac{d}{dx_2}(x_1^2 + x_1 x_2^3)\Big|_{x=a} = 2x_1 x_2^2|_{x=a} \\ &= 2a_1 a_2^2 = 4 + 27 = 72\end{aligned}$$

APPENDIX C. ABOUT JACOBIAN

Note 113. Consider a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$.

Note 114. The Jacobian of f at $w \in \mathbb{R}^n$, is denoted as $J_w(f)$ and it is an $m \times n$ matrix with (i, j) elements such as

$$[J_w(f)]_{i,j} = \frac{d}{dw_j} f_i(w)$$

for $i = 1, \dots, m$ and $j = 1, \dots, n$.

Note 115. To align with standard notation in statistics, we will use the notation

$$\frac{df}{dw}, \text{ or } \frac{d}{dw} f(w)$$

for $J_w(f)$.

Note 116. Some properties

- Let functions $f(w) = Aw$, for matrix $A \in \mathbb{R}^{m,n}$ and vector $w \in \mathbb{R}^n$. Then $\frac{df}{dw}(w) = A$.
- Let functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^k \rightarrow \mathbb{R}^n$. The Jacobian of the composition function $f \circ g : \mathbb{R}^k \rightarrow \mathbb{R}^m$ with $f(w) = f(g(w))$ at w is

$$\frac{d}{d\xi} f \circ g(\xi)\Big|_{\xi=w} = \frac{df}{dg}(g)\Big|_{g=g(w)} \frac{dg}{d\xi}(\xi)\Big|_{\xi=w}$$

or more compactly

$$\frac{d}{dw} f \circ g = \frac{df}{dg} \frac{dg}{dw}$$

Example 117. Let $g : \mathbb{R}^k \rightarrow \mathbb{R}^n$ with $g(w) = Aw$ where matrix $A \in \mathbb{R}^{n,k}$. Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$. Compute $\frac{d}{dw} \sigma \circ g(w)$.

Hint: Adopt notation $\sigma(a) = (\sigma(a_1), \dots, \sigma(a_n))$ when $a \in \mathbb{R}^d$.

Solution. It is

$$\begin{aligned}\frac{d}{dw} \sigma \circ g(w) &= \frac{d}{dg} \sigma(g) \Big|_{g=g(w)} \frac{d}{dw} g(w) \\ &= \frac{d}{dg} \sigma(g) \Big|_{g=Aw} \frac{d}{dw} g(w) \\ &= \text{diag}(\sigma(Aw)) A\end{aligned}$$

This is because

$$\left[\frac{d}{dw} g(w) \right]_{i,j} = \left[\frac{d}{dw} Aw \right]_{i,j} = \frac{\partial}{\partial w_j} \sum_{k=1}^k A_{i,k} w_k = A_{i,j}$$

and because

$$\left[\frac{d}{dg} \sigma(g) \right]_{i,j} = \left[\frac{d(\sigma(g_1), \dots, \sigma(g_n))}{d(g_1, \dots, g_n)} \right]_{i,j} = \frac{\partial}{\partial g_j} \sigma(g_i) = \begin{cases} \sigma(g_i), & i = j \\ 0, & i \neq j \end{cases}$$

Handout 6: Support Vector Machines

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce the Support Vector Machines as a procedure. Motivation, set-up, description, computation, and implementation. We focus on the classical treatment.

Reading list & references:

- (1) Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
 - Ch. 15 (pp. 167-170, 171-172, 176-177) Support Vector Machine
- (2) Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.
 - Ch. 7.1 Sparse Kernel Machines/Maximum marginal classifiers

1. INTRO AND MOTIVATION

Note 1. Support Vector Machines (SVM) is a ML procedure for learning linear predictors in high-dimensional feature spaces with regards the sample complexity challenges. Due to a duality property, SVM have sparse solutions, so that predictions for new inputs depend only on quantities evaluated at a subset of the training data points.

Definition 2. Let $w \neq 0$. Hyperplane in space $\mathcal{X} \subseteq \mathbb{R}^d$ is called the sub-set

$$S = \left\{ x \in \mathbb{R}^d : \langle w, x \rangle + b = 0 \right\}.$$

It separates \mathcal{X} in two half-spaces

$$S_+ = \left\{ x \in \mathbb{R}^d : \langle w, x \rangle + b > 0 \right\}$$

and

$$S_- = \left\{ x \in \mathbb{R}^d : \langle w, x \rangle + b < 0 \right\}$$

Definition 3. Halfspace (hypothesis space) is hypotheses class \mathcal{H} designed for binary classification problems, $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$ defined as

$$\mathcal{H} = \left\{ x \mapsto \text{sign}(\langle w, x \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R} \right\},$$

where b is called bias.

Definition 4. Each halfspace hypothesis $h \in \mathcal{H}$ has form

$$(1.1) \quad h_{w,b}(x) = \text{sign}(\langle w, x \rangle + b)$$

, it takes an input in $\mathcal{X} \subseteq \mathbb{R}^d$ and returns an output in $\mathcal{Y} = \{-1, +1\}$. We may refer to it as halfspace (w, b) as this setting determines it.

Note 5. Let $S = \{(x_i, y_i)\}_{i=1}^m$ be a training set of examples with $x_i \in \mathbb{R}^d$ the features and $y_i \in \{-1, +1\}$ the labels.

Definition 6. The training set S is **linearly separable** if there exists a halfspace (w, b) such that for all $i = 1, \dots, n$

$$y_i = \text{sign}(\langle w, x_i \rangle + b)$$

or equivalently

$$y_i (\langle w, x_i \rangle + b) > 0$$

Note 7. Let the loss be $\ell((w, b), z) = 1(y_i \neq \text{sign}(\langle w, x_i \rangle + b))$, and hence the Empirical Risk Function be $R_S(w, b) = \frac{1}{m} \sum_{i=1}^m \ell((w, b), z_i)$. The Empirical Risk Minimisation (ERM) halfspace (w^*, b^*) is

$$(w^*, b^*) = \arg \min_{w, b} (R_S(w, b)) = \arg \min_{w, b} \left(\frac{1}{m} \sum_{i=1}^m \ell((w, b), z_i) \right)$$

Definition 8. Margin of a hyper-plane with respect to a training set is defined to be the minimal distance between a point in the training set and the hyper-plane.

Note 9. Support Vector Machines (SVM) aims at learning the maximum margin separating hyperplane Figure (1.1; Right). The rational is that if a hyperplane has a large margin, then it will still separate the training set even if we slightly perturb each instance.

Example 10. Figure (1.1; Left) shows two different separating hyper-planes for the same data set, Figure (1.1; Right) shows the maximum margin hyper-plane: the margin γ is the distance from the hyper-plane (solid line) to the closest points in either class (which touch the parallel dotted lines). It is reasonable to prefer as a predictive rule the hyperplane on the right.

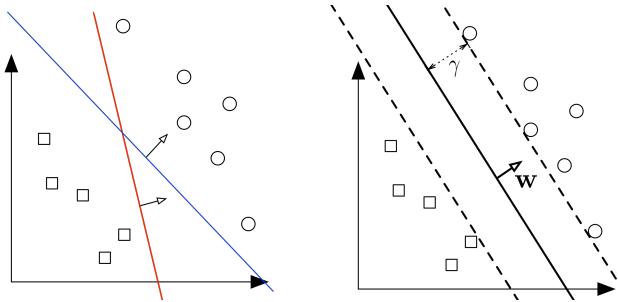


FIGURE 1.1

2. HARD SUPPORT VECTOR MACHINE

Note 11. Hard Support Vector Machine (Hard-SVM) is the learning rule in which we return an ERM hyperplane that separates the training set with the largest possible margin.

Assumption 12. Assume the training sample $S = \{(x_i, y_i)\}_{i=1}^m$ is linearly separable.

Algorithm 13. (Hard-SVM) Given a linearly separable training sample $S = \{(x_i, y_i)\}_{i=1}^m$ the Hard-SVM rule for the binary classification problem is:

*Solve*¹

$$(2.1) \quad (\tilde{w}, \tilde{b}) = \arg \min_{(w,b)} \|w\|_2^2$$

$$(2.2) \quad \text{subject to: } y_i (\langle w, x_i \rangle + b) \geq 1, \quad \forall i = 1, \dots, m$$

Scale

$$\hat{w} = \frac{\tilde{w}}{\|\tilde{w}\|}, \quad \text{and} \quad \hat{b} = \frac{\tilde{b}}{\|\tilde{w}\|}$$

Note 14. Following we show why Algorithm 13 produces a Hard-SVM hyperplane stated in Note 11.

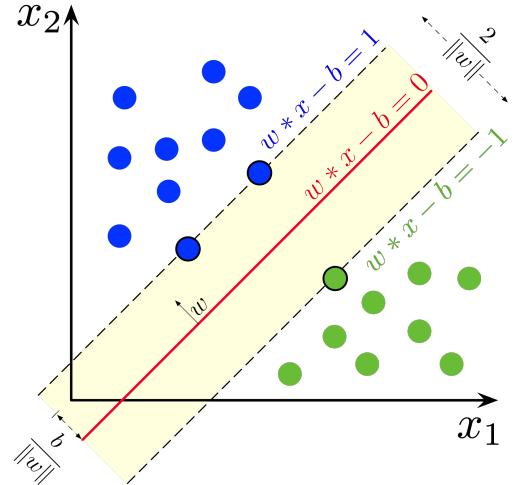
Fact 15. *The distance between a point x and the hyperplane defined by (w, b) with $\|w\| = 1$ is $|\langle w, x \rangle + b|$.*

Proof. We skip it. □

Note 16. On the right, see the geometry of Algorithm 13.

Note 17. Hard-SVM selects two parallel hyperplanes that separate the two classes of data so that the distance between them is as large as possible. The predictive hyperplane (rule) is the hyperplane that lies halfway between them.

Note 18. Hard-SVM in Algorithm 13 searches for the hyperplane with minimum norm w among all those that separate the data and have distance greater or equal to 1.



Proof. (Sketch of the proof of Algorithm 13)

- (1) Based on Note 11, and Fact 15, the closest point in the training set to the separating hyperplane has distance

$$\min_i (|\langle w, x_i \rangle + b|)$$

hence, by definition, the Hard-SVM hypothesis should be such as

$$(2.3) \quad (w^*, b^*) = \arg \max_{(w,b): \|w\|=1} \left(\min_i (|\langle w, x_i \rangle + b|) \right)$$

$$(2.4) \quad \text{subject to } y_i (\langle w, x_i \rangle + b) > 0, \quad \forall i = 1, \dots, m$$

- (2) If there is a solution in (2.3) then (2.3) is equivalent to

$$(2.5) \quad (w^*, b^*) = \arg \max_{(w,b): \|w\|=1} \left(\min_i (y_i (\langle w, x_i \rangle + b)) \right)$$

¹It is a quadratic programming problem.

(3) Next we show that 2.5 is equivalent to the output of Algorithm 13; i.e. $(w^*, b^*) = (\hat{w}, \hat{b})$.

Let $\gamma^* := \min_i (|\langle w^*, x_i \rangle + b^*|)$. Firstly, because

$$y_i (\langle w^*, x_i \rangle + b^*) \geq \gamma^* \Leftrightarrow y_i \left(\frac{\langle w^*, x_i \rangle + b^*}{\gamma^*} \right) \geq 1$$

$\left(\frac{w^*}{\gamma^*}, \frac{b^*}{\gamma^*} \right)$ satisfies condition (2.2). Secondly, I have $\|w_0\| \leq \left\| \frac{w^*}{\gamma^*} \right\| = \frac{1}{\gamma^*}$ because of (2.1) and because of $\|w^*\| = 1$. Hence, for all $i = 1, \dots, m$, it is

$$y_i \left(\langle \hat{w}, x_i \rangle + \hat{b} \right) = \frac{1}{\|w^*\|} y_i (\langle w_0, x_i \rangle + b_0) \geq \frac{1}{\|w^*\|} \geq \gamma^*$$

Hence (\hat{w}, \hat{b}) is the optimal solution of (2.5).

□

Definition 19. **Homogeneous halfspaces** in SVM is the case where the halfspaces pass from the origin; that is when the bias term in 2.2 is zero $b = 0$.

3. SOFT SUPPORT VECTOR MACHINE

Note 20. Hard-SVM assumes the strong Assumption 12 that the training set is linearly separable, that might not always be the case, and hence there is need to derive a procedure that weakens this assumption.

Note 21. Soft Support Vector Machine (Soft-SVM) aims to relax the strong assumption of Hard-SVM that the training set is linearly separable (2.4) with purpose to extend the scope of application. Soft-SVM is given below. I.e., Soft-SVM does not assume Assumption 12.

Algorithm 22. (*Soft-SVM*) Given a training sample $S = \{(x_i, y_i)\}_{i=1}^m$ the Soft-SVM rule for the binary classification problem is:

Solve

$$(3.1) \quad (w^*, b^*, \xi^*) = \arg \min_{(w, b, \xi)} \left(\lambda \|w\|_2^2 + \frac{1}{m} \sum_{i=1}^m \xi_i \right)$$

$$(3.2) \quad \text{subject to: } y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad \forall i = 1, \dots, m$$

$$(3.3) \quad \xi_i \geq 0, \quad \forall i = 1, \dots, m$$

Note 23. To relax the linearly separable training set assumption, Soft-SVM relies on replacing the “harder” constraint (2.2) with the “softer” one in 3.2 through the introduction of non-negative unknown quantities $\{\xi_i\}_{i=1}^m$ controlling how much the separability assumption (2.2) is violated. Soft-SVM learns all (w, b, ξ) via the minimization part in (3.1) where the trade off between the two terms is controlled via the user specified parameter λ .

Proposition 24. Consider the hinge loss function

$$\ell((w, b), z) = \max(0, 1 - y(\langle w, x \rangle + b))$$

and hence the Empirical Risk Function

$$R_S((w, b)) = \frac{1}{m} \sum_{i=1}^m \max(0, 1 - y_i(\langle w, x_i \rangle + b))$$

Then the solution of Algorithm 22 is equivalent to the regularization problem

$$(w^*, b^*) = \arg \min_{(w, b)} \left(R_S((w, b)) + \lambda \|w\|_2^2 \right)$$

Proof. In Algorithm 22, we consider

$$(3.4) \quad \arg \min_{(w, b)} \left(\min_{\xi} \left(\lambda \|w\|_2^2 + \frac{1}{m} \sum_{i=1}^m \xi_i \right) \right)$$

Consider (w, b) fixed and focus on the inside minimization. From (3.2), it is $\xi_i \geq 1 - y_i(\langle w^*, x_i \rangle + b^*)$, and from (3.3), it is $\xi_i \geq 0$. If $y_i(\langle w, x_i \rangle + b) \geq 1$, the best assignment in 3.4 is $\xi_i = 0$ because it is $\xi_i \geq 0$ from (3.3) and I need to minimize (3.4) wrt ξ_i 's. If $y_i(\langle w, x_i \rangle + b) \leq 1$, the best assignment in (3.4) is $\xi_i = 1 - y_i(\langle w, x_i \rangle + b)$ because I need to minimize w.r.t ξ . Hence $\xi_i = \max(0, 1 - y_i(\langle w, x_i \rangle + b))$. \square

Note 25. Hence the Soft-SVM is a binary classification problem with hinge loss function and regularization term biasing toward low norm separators.

Note 26. Given Proposition 24, Soft-SVM in Algorithm 22 can be learned via any variation of SGD, eg online SGD (batch size $m = 1$) with recursion

$$\varpi^{(t+1)} = \varpi^{(t)} - \eta_t v_t$$

$$\text{where } v_t = \begin{cases} y^{(t)} \langle \varpi^{(t)}, \chi^{(t)} \rangle & \text{if } y^{(t)} \langle \varpi, \chi^{(t)} \rangle \geq 1 \\ -y^{(t)} \chi^{(t)} & \text{otherwise} \end{cases}, \varpi = (b^{(t)}, w^{(t)})^\top \text{ and } \chi = (1, x^{(t)})^\top.$$

4. SUPPORT VECTORS

Lemma 27. (*Fritz John optimality conditions*) Suppose that

$$\begin{aligned} w^* &= \arg \min_w f(w) \\ \text{s.t. } g_i(w) &\leq 0, \quad \forall i = 1, \dots, m \\ h_j(x) &= 0, \quad \forall j = 1, \dots, n \end{aligned}$$

where f, g_1, \dots, g_m are differentiable. Then there exists $\alpha_i \geq 0$ for $i = 1, \dots, m$ and $\beta_j \in \mathbb{R}$ such that

$$\nabla_w f(w^*) + \sum_{i \in I} \alpha_i \nabla_w g_i(w^*) + \sum_{j \in \mathcal{J}} \beta_j \nabla_w h_j(w^*) = 0$$

where $I = \{i : g_i(w^*) = 0\}$.

Theorem 28. Let \hat{w} such that

$$(4.1) \quad \hat{w} = \arg \min_w \|w\|_2^2$$

$$(4.2) \quad \text{subject to: } y_i \langle w, x_i \rangle \geq 1, \forall i = 1, \dots, m$$

then there exists coefficients $\alpha_i \in \mathbb{R}$ for $i = 1, \dots, m$ such that ²

$$\hat{w} = \sum_{i \in I} \alpha_i x_i$$

where $I = \{i : \langle \hat{w}, x_i \rangle = 1\}$.

Definition 29. Support vectors are called the examples $\{x_i : i \in I\}$ in the training data set.

Note 30. The support vectors have distance $1/\|\hat{w}\|$ from the separating hyperplane. See Figure 4.1.

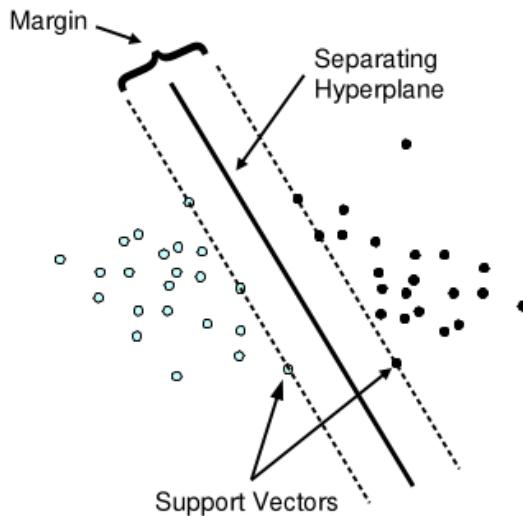


FIGURE 4.1

5. DUALITY

Fact 31. (Karush–Kuhn–Tucker (KKT) conditions) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an objective convex function, let $\{g_i : \mathbb{R}^n \rightarrow \mathbb{R}\}_{i=1}^m$ inequality constraint convex functions, and let $\{h_j : \mathbb{R}^n \rightarrow \mathbb{R}\}_{j=1}^n$ equality constraint functions.

Primal problem: Consider the following convex nonlinear minimization problem, with solution x^*

$$(5.1) \quad \begin{aligned} p^* &= \underset{x}{\text{minimize}} (f(x)) \\ &\text{s.t. } g_i(x) \leq 0, \quad \forall i = 1, \dots, m \\ &\quad h_j(x) = 0, \quad \forall j = 1, \dots, n \end{aligned}$$

² y_i is absorbed in α_i 's
Page 6

Lagrangian function: To the problem we associate the Lagrangian $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^n$ with

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^m \alpha_i g_i(x) + \sum_{j=1}^n \beta_j h_j(x)$$

Lagrangian dual problem: The associated Lagrangian dual problem is

$$\begin{aligned} d^* &= \underset{\alpha}{\text{maximize}} \left(\min_x (L(x, \alpha, \beta)) \right) \\ &\text{s.t. } \alpha_i \geq 0 \\ &\quad \beta_j \in \mathbb{R} \end{aligned}$$

where

$$\tilde{L}(\alpha, \beta) := \min_x (L(x, \alpha, \beta))$$

is called dual function

(Weak duality): In general it is $p^* \geq d^*$.

(Strong duality via Slater condition): If the primal problem (5.1) is convex, and satisfies the weak Slater's condition, i.e.

$$(\exists x_0 \in \mathcal{D}) : (g_i(x_0) < 0, \forall i = 1, \dots, n) \text{ and } (h_i(x_0) = 0, \forall i = 1, \dots, n)$$

then strong duality holds, that is: $p^* = d^*$. It other words

$$\min_x \max_{\alpha \geq 0, \beta} (L(x, \alpha, \beta)) = \max_{\alpha \geq 0, \beta} \min_x (L(x, \alpha, \beta))$$

Solution: With an optimal vector x^* in (5.1) there is associated a non-negative vector α^* and vector β^* such that

$$0 = \nabla_{(x, \alpha, \beta)} L(x^*, \alpha^*, \beta^*)$$

KKT conditions: Karush–Kuhn–Tucker (KKT) conditions are necessary and sufficient conditions for x^* to be local minimum

$$\begin{aligned} 0 &= \nabla f(x^*) + \sum_{j=1}^n \beta_j \nabla h_j(x^*) + \sum_{i=1}^m \alpha_i \nabla g_i(x^*) && \text{Stationarity} \\ g_i(x^*) &\leq 0, \quad \forall i = 1, \dots, m && \text{Primal feasibility} \\ h_j(x^*) &= 0, \quad \forall j = 1, \dots, n \\ \alpha_i &\geq 0 \quad \forall i = 1, \dots, m && \text{Dual feasibility} \\ \alpha_i g_i(x^*) &= 0, \quad \forall i = 1, \dots, m && \text{Complementary slackness} \end{aligned}$$

5.1. Hand SVM.

Note 32. The problem in Hard-SVM Algorithm 13 (2.2) and (2.2), can be re-written in the form

$$(5.3) \quad \min_w \left(\frac{1}{2} \|w\|_2^2 + g(w) \right)$$

for

$$g(w) = \max_{\alpha \in \mathbb{R}^m: \alpha \geq 0} \sum_{i=1}^m \alpha_i (1 - y_i (\langle w, x_i \rangle + b)) = \begin{cases} 0 & \text{if } y_i (\langle w, x_i \rangle + b) \geq 1 \\ \infty & \text{else} \end{cases}$$

Hence (5.3) is equivalent to

$$\begin{aligned} (5.4) \quad & \min_w \max_{\alpha \in \mathbb{R}^m: \alpha \geq 0} \left(\frac{1}{2} \|w\|_2^2 + \sum_{i=1}^m \alpha_i (1 - y_i (\langle w, x_i \rangle + b)) \right) \\ & \geq \max_{\alpha \in \mathbb{R}^m: \alpha \geq 0} \min_{w, b} \underbrace{\left(\frac{1}{2} \|w\|_2^2 + \sum_{i=1}^m \alpha_i (1 - y_i (\langle w, x_i \rangle + b)) \right)}_{=L(w, b; \alpha)} \end{aligned}$$

where (5.4) is called the weak duality. Strong duality is when the equality hold. In our case the equality holds, but we do not go into details. Here, keeping α fixed, $L(w, \alpha, b)$ is minimized when

$$(5.5) \quad 0 = \nabla_w L(w, \alpha, b)|_{(w^*, b^*)} \implies w^* = \sum_{i=1}^m \alpha_i y_i x_i$$

$$(5.6) \quad 0 = \nabla_b L(w, \alpha, b)|_{(w^*, b^*)} \implies 0 = \sum_{i=1}^m \alpha_i y_i$$

The dual function is

$$\begin{aligned} \tilde{L}(\alpha) &= \min_{w, b} (L(w, \alpha, b)) = \frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i x_i \right\|_2^2 - \sum_{i=1}^m \alpha_i \left(y_i \left(\sum_{j=1}^m \alpha_j y_j x_j, x_i \right) + b \right) - 1 \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle x_j, x_i \rangle \end{aligned}$$

Then dual problem is

$$\max_{\alpha \in \mathbb{R}^m: \alpha \geq 0} \left(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle x_j, x_i \rangle \right)$$

Note 33. The Dual problem of the (Primal) problem in Hard-SVM Algorithm 13 is

$$\begin{aligned} (5.7) \quad \alpha^* &= \arg \max_{\alpha \in \mathbb{R}^m: \alpha \geq 0} \left(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle x_j, x_i \rangle \right) \\ &\text{subject to } 0 = \sum_{i=1}^m \alpha_i y_i \end{aligned}$$

Note 34. Once the optimal α^* are computed from (5.7), the optimal weights w^* can be computed (from (5.5)) as

$$(5.8) \quad w^* = \sum_{i=1}^m \alpha_i^* y_i x_i$$

Note 35. If $\alpha_i^* = 0$, the example (x_i, y_i) does not contribute to (5.8). If $\alpha_i^* \neq 0$, (x_i, y_i) contributes to (5.8). Moreover if $\alpha_i^* \neq 0$ KKT condition (5.2)

$$\alpha_i^* (y_i (\langle w^*, x_i \rangle + b^*) - 1) = 0$$

implies that $y_i (\langle w^*, x_i \rangle + b^*) = 1$ meaning that x_i is on the boundary of the margin. Features x_i associated to $\alpha_i^* \neq 0$ which contribute to the evaluation of the optimal weights w^* and hence the evaluation of the separating predictive rule $h_{w,b}$ are called **supporting vectors** (see Figure 4.1).

Note 36. Given optimal α^* , and letting $\mathcal{I} = \{i : y_i (\langle w^*, x_i \rangle + b^*) - 1 = 0\}$, 5.8 becomes

$$(5.9) \quad w^* = \sum_{i \in \mathcal{I}} \alpha_i^* y_i x_i$$

Note 37. The bias can be computed by KKT condition (5.2), it is

$$\alpha_i^* (y_i (\langle w^*, x_i \rangle + b^*) - 1) = 0$$

Consider $\mathcal{I} = \{i : y_i (\langle w^*, x_i \rangle + b^*) - 1 = 0\}$ then, for $i \in \mathcal{I}$ it is

$$y_i^2 (\langle w^*, x_i \rangle + b^*) - y_i = 0 \stackrel{y_i^2 = 1}{\implies} \langle w^*, x_i \rangle + b^* - y_i = 0$$

and summing up all $i \in \mathcal{I}$, I get

$$(5.10) \quad \sum_{i \in \mathcal{I}} (\langle w^*, x_i \rangle + b^* - y_i) = 0 \implies b^* = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (y_i - \langle w^*, x_i \rangle)$$

Note 38. Therefore the predictive halfspace hypothesis is given as

$$(5.11) \quad \begin{aligned} h_{w,b}(x) &= \text{sign}(\langle w^*, x \rangle + b^*) \\ &= \text{sign}\left(\sum_{i \in \mathcal{I}} \alpha_i^* y_i \langle x_i, x \rangle + b^*\right) \end{aligned}$$

with $\mathcal{I} = \{i : y_i (\langle w^*, x_i \rangle + b^*) - 1 = 0\}$.

Note 39. Compared to the Primal problem (Algorithm 13), the dual problem (Note 33) is computationally desirable in cases that the training data size m is smaller than the number of the dimensionality of the feature space d ; i.e. $d \gg m$. The solution of the Primal quadratic programming problem in $d+1$ and complexity $O(d+1)$, while the dual quadratic programming problem has m variables and complexity $O(m)$.

Note 40. Compared to the Primal problem, the dual problem (Note 33), can provide sparse solutions relying on a small number of support vectors (see 5.9, 5.10, and 5.11).

Note 41. The importance of the existence of the dual problem is that eg 5.7 and (5.11) involves the inner products between instances and does not require the direct access to specific elements of features within instance. For instance, if we consider the hypothesis (1.1) as an expansion of bases $h_{w,b}(x) = \text{sign}(\langle w, \phi(x) \rangle + b)$ with $\phi(x) = (\phi_1(x), \dots, \phi_d(x))$ with a large d such as $d \gg m$ then,

based on (5.11), the separation rule is

$$h_{w,b}(x) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i k(x', x) + b \right)$$

and 5.7 becomes

$$\alpha^* = \arg \max_{\alpha \in \mathbb{R}^m : \alpha \geq 0} \left(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j k(x', x) \right)$$

with $k(x', x) = \langle \phi(x'), \phi(x) \rangle$. As we will see in the “kernel methods” lecture, in specific case, one can specify the “kernel” function $k(\cdot, \cdot)$ (having with specific desirable properties) avoiding the direct specification of a possibly high-dimensional dictionary of features $\{\phi_j(\cdot)\}$.

Note 42. In the Soft-SVM case, the solution is the same as in the Hard-SVM (5.8) and (5.11), the only difference is that in the quadratic programming problem (5.7) it is subject to $\alpha_j \in [0, C]$ where $C = 1/2\lambda$, it is called “cost” and controls the cost of having the constraint violation by adding the ξ'_i s.

APPENDIX A. RECALL

The following is part of “Handout 1: Elements of convex learning problems”

Definition 43. Convex learning problem is a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ that the hypothesis class \mathcal{H} is a convex set, and the loss function ℓ is a convex function for each example $z \in \mathcal{Z}$.

Definition 44. Convex-Lipschitz-Bounded Learning Problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with parameters ρ , and B , is called the learning problem whose the hypothesis class \mathcal{H} is a convex set, for all $w \in \mathcal{H}$ it is $\|w\| \leq B$, and the loss function $\ell(\cdot, z)$ is convex and ρ -Lischitz function for all $z \in \mathcal{Z}$.

Definition 45. Convex-Smooth-Bounded Learning Problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with parameters β , and B , is called the learning problem whose the hypothesis class \mathcal{H} is a convex set, for all $w \in \mathcal{H}$ it is $\|w\| \leq B$, and the loss function $\ell(\cdot, z)$ is convex, nonnegative, and β -smooth function for all $z \in \mathcal{Z}$.

Handout 8: Gaussian process regression

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce the ideas of learning machines by introducing data into high-dimensional feature spaces for accuracy gains; introduce the kernel trick, and kernel functions.

Reading list & references:

- (1) Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
 - Ch. 16.2 Support Vector Machine
- (2) Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.
 - Ch. 6.1, 6.2 Kernel methods

1. INTRO AND MOTIVATION

Note 1. Consider the Soft SVM with predictive rule $h(x) = \text{sign}(\eta(x))$ with separator $\eta(x) = w_1x_1 + w_2x_2 + b = \langle w, x \rangle + b$ and $x \in \mathbb{R}^2$. It can address learning problems where the data can (up to some degree of violation) be separated by a line (Figure 1.1a). In more challenging cases where the geometry is strongly non-linear this can totally fail (Figure 1.1b).

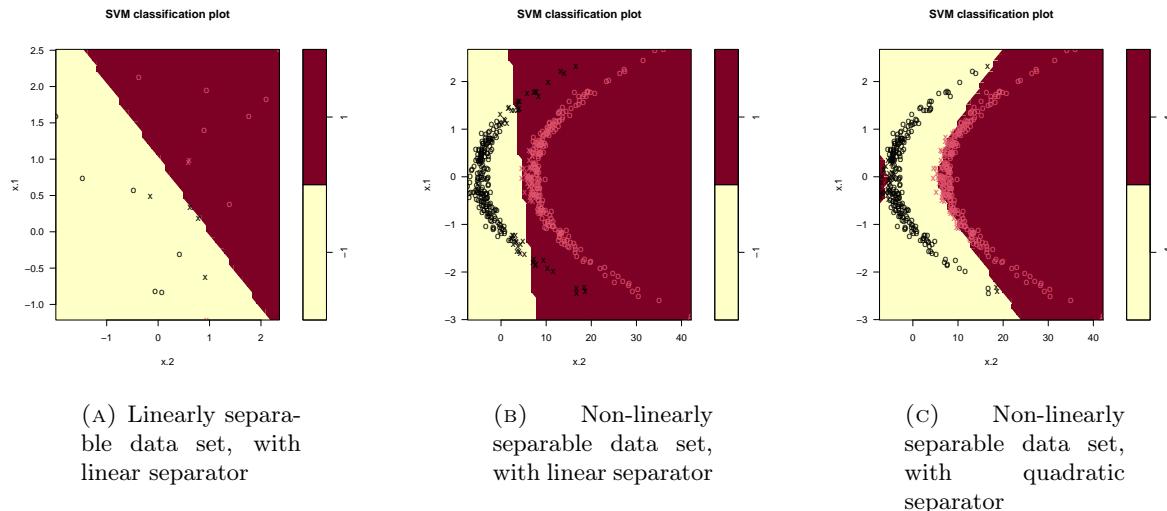


FIGURE 1.1. Soft SVM from Computer practical 4

Note 2. The accuracy of predictive rule can be improved if I take into account the curvature by adding a quadratic term in the 2nd dimension as $\eta'(x) = w'_1x_1 + w'_2x_2 + w'_3x_3 + b = \langle w', \psi(x) \rangle + b$ where $\psi(x) = (x_1, x_2, x_2^2)$ and learning the w' . It works, (Figure 1.1c).

Note 3. Consequently, in order to improve the expressiveness of a hypothesis class $\mathcal{H} = \{x \mapsto f(\langle w, x \rangle)\}$ (for some function f) with purpose to learn a more accurate predictive rule, it is reasonable to consider an embedding $\psi(x)$, which can possibly be a vector of bases functions such as polynomials, splines, etc..., and work on the learning problem with $\mathcal{H} = \{x \mapsto f(\langle w, \psi(x) \rangle)\}$.

Note 4. The above may drastically increase the dimensionality of the problem hence the computational cost and required number of example to train. This challenge is addressed by the Kernel trick.

Note 5. In general, kernel trick allows the design of powerful and cheap extensions of many well known algorithms.

2. PROJECTIONS IN FEATURE SPACES (IMPROVING EXPRESSIVE POWER)

Note 6. To make the class of hypotheses more expressive with purpose to improve accuracy, we can first map the original instance space $x \in \mathcal{X}$ into another space \mathcal{F} (possibly of a higher dimension) via an embedding ψ and then learn a hypothesis in that space.

Summary 7. Consider a hypothesis class $\mathcal{H} = \{x \mapsto \langle w, x \rangle : w \in \mathbb{R}^n\}$ where the predictive rule $h \in \mathcal{H}$ defined over \mathcal{X} is to be trained against data set $\mathcal{S} = \{z_i = (x_i, y_i)\}_{i=1}^m$. The basic paradigm involves:

- (1) Choose a mapping $\psi : \mathcal{X} \rightarrow \mathcal{F}$ with $\psi(x) := (\psi_1(x), \dots, \psi_d(x))$ for some feature space \mathcal{F} .
- (2) Create the image sequence $\tilde{\mathcal{S}} = \left\{ z_i^\psi = (\psi(x_i), y_i) \right\}_{i=1}^m$ from the original training set \mathcal{S} .
- (3) Train a linear predictor h against $\tilde{\mathcal{S}}$.
- (4) Predict the label or the output of a new point x^{new} by $h^\psi(x^{\text{new}}) := h \circ \psi(x^{\text{new}}) = h(\psi(x^{\text{new}}))$

Note 8. The introduction of mapping $\psi : \mathcal{X} \rightarrow \mathcal{F}$ induces

- (1) probability distribution G^ψ over domain $\mathcal{X} \times \mathcal{F}$ with $G^\psi(A) = G(\psi^{-1}(A))$ for every set $A \subseteq \mathcal{X} \times \mathcal{F}$.
- (2) predictive rule $h^\psi(\cdot) := h \circ \psi(\cdot)$, where $h \circ \psi(\cdot) = h(\psi(\cdot))$
- (3) risk function $R_{G^\psi}(h) := R_G(h \circ \psi)$, as

$$R_G(h \circ \psi) = \int \ell(h \circ \psi, z = (x, y)) dG(z) = \int \ell(h, z^\psi) dG^\psi(x, y) = R_{G^\psi}(h)$$

Definition 9. A Hilbert space is a vector space, with an inner product, which is also complete¹.

Lemma 10. If \mathcal{X} is a linear subspace of a Hilbert space, then every $x \in \mathcal{X}$ can be written as $x = u + v$ where $u \in \mathcal{X}$ and $\langle u, v \rangle = 0$ for all $v \in \mathcal{X}$.

¹Going deeper to completeness is out of the scope.

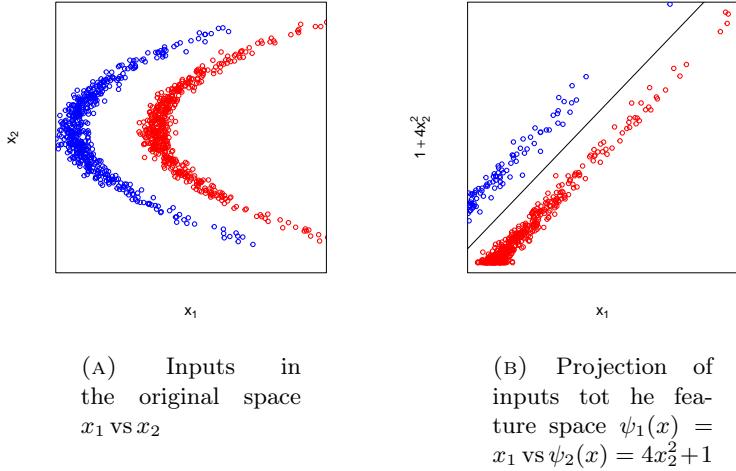


FIGURE 2.1. Projection of the inputs living in the original space to the feature space

Note 11. Feature space \mathcal{F} is a Hilbert space preferably due to Lemma 10 that enables the Kernel trick via the representation Theorem 21. Eg, a Euclidean space such as \mathbb{R}^d for some d . That includes infinite dimensional spaces.

Note 12. As feature mapping ψ any function that maps the original instances \mathcal{X} into some Hilbert space \mathcal{F} can be used.

Note 13. The success of the learning paradigm in Summary 7 depends on choosing a good embedding ψ for the particular learning task. Eg, in SVM, ψ will make the image of the data distribution (close to being) linearly separable in the feature space \mathcal{F} , thus making the resulting learning algorithm a good learner for a given task (Figure 2.1). This requires prior knowledge of the problem (In Section 4, we see popular recipes for that).

Note 14. Using a $\psi(x) := (\psi_1(x), \dots, \psi_d(x))$ that is high dimensional (d is too large) may improve accuracy (expressiveness) of the learner (e.g. recall in polynomial regression increasing the polynomial degree). However this increases the computational effort/cost required to perform calculations to minimize the associated risk function in the high dimensional space, as well as we need more data. This is addressed via the Kernel trick.

Note 15. Recall the feed forward neural network (let's say with 1 hidden layer) formula mapping x to $h(x)$:

$$h(x) = \sigma_2 \left(\sum_{j=1}^c w_{2,1,j} \sigma_1 \left(\sum_{i=1}^n w_{1,j,i} x_i \right) \right).$$

It can be considered as $h(x) = \sigma_2(\langle w_2, \psi(x) \rangle)$ with where the output of the hidden layer is the embedding $\psi(x) = (\sigma_1(\sum_{i=1}^n w_{1,1,i} x_i), \dots, \sigma_1(\sum_{i=1}^n w_{1,c,i} x_i))$. Interestingly, FFNN, can be considered as providing an adaptive (semi-parametric) ψ as the weights $\{w_{1,j,i}\}$ are adapted during the training. Hence, FFNN performs a deformation of the input space in the same spirit to the projections

in feature spaces method. Adding more hidden layers (multi-perceptron) potentially can improve this.

3. THE KERNEL TRICK

Definition 16. Kernel function K is defined as $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ with $K(x, x') = \langle \psi(x), \psi(x') \rangle$ given an embedding $\psi(x)$ of some domain space \mathcal{X} into some Hilbert space \mathcal{F} . Kernel functions describe inner products in the feature space \mathcal{F} .

Note 17. Kernel function K can be considered as specifying similarity between instances via the embedding $\psi(\cdot)$ as mapping the domain set \mathcal{X} into a space where these similarities are realized as inner products.

Problem 18. (Learning problem) Consider a prediction rule $h : \mathcal{X} \rightarrow \mathcal{Y}$ with $h(x) = \langle w, \psi(x) \rangle$ which is trained against a training sample $\{z_i = (x_i, y_i)\}_{i=1}^m$ with the following general optimization problem

$$(3.1) \quad \underset{w}{\text{minimize}}(f(\langle w, \psi(x_1) \rangle, \dots, \langle w, \psi(x_m) \rangle) + R(\|w\|)),$$

where $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is an arbitrary function and $R : \mathbb{R}_+ \rightarrow \mathbb{R}$ is a monotonically non-decreasing function.

Example 19. In Soft SVM (Proposition 24, Handout 6), it is $f(a_1, \dots, a_m) = \frac{1}{m} \sum_{i=1}^m \max\{0, 1 - y_i a_i\}$ and $R(a) = \lambda a^2$.

Note 20. The following result states a duality in the learning problem 18 that facilitates the implementation of the extension to a possibly high dimensional feature space (hence improving the expressiveness/accuracy) by using kernel functions (hence reducing dimensionality, computational cost, and required data size).

Theorem 21. (*Representation theorem*) Assume mapping $\psi : \mathcal{X} \rightarrow \mathcal{F}$ where \mathcal{F} is a Hilbert space. There exists a vector $\alpha \in \mathbb{R}^m$ such that $w = \sum_{i=1}^m \alpha_i \psi(x_i)$ is the optimal solution of (3.1) in Problem 18.

Proof. Let w^* be the optimal solution of (3.1). Because w^* is element of Hilbert space, it can be written as $w^* = \sum_{i=1}^m \alpha_i \psi(x_i) + u$ where $\langle u, \psi(x_i) \rangle = 0$ for all $i = 1, \dots, m$. Set $w := w^* - u$.

Because $\|w^*\|^2 = \|w\|^2 + \|u\|^2$ it is $\|w\| \leq \|w^*\|$ implying that

$$R(\|w\|) \leq R(\|w^*\|).$$

Because $\langle w, \psi(x_i) \rangle = \langle w^* - u, \psi(x_i) \rangle = \langle w^*, \psi(x_i) \rangle$ for all $i = 1, \dots, m$, it is

$$f(\langle w, \psi(x_1) \rangle, \dots, \langle w, \psi(x_m) \rangle) = f(\langle w^*, \psi(x_1) \rangle, \dots, \langle w^*, \psi(x_m) \rangle)$$

Then the objective function of (3.1) at w is less than or equal to that of the minimizer w^* which implies that $w = \sum_{i=1}^m \alpha_i \psi(x_i)$ is an optimal solution. \square

Note 22. Let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a kernel function with $K(x, x') = \langle \psi(x), \psi(x') \rangle$. According to the representation Theorem 21, the Learning problem 18, can be equivalently addressed by re-writing the learning predictive rule as

$$h_\alpha(x) = \sum_{i=1}^m \alpha_i K(x_i, x)$$

and learning $\{\alpha_i\}$ as the solutions of

$$(3.2) \quad \underset{\alpha}{\text{minimize}} \left(f \left(\sum_{i=1}^m \alpha_i K(x_i, x), \dots, \sum_{i=1}^m \alpha_i K(x_m, x) \right) + R \left(\sqrt{\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(x_i, x_j)} \right) \right),$$

This is because

$$\langle w, \psi(x_j) \rangle = \left\langle \sum_{i=1}^m \alpha_i \psi(x_i), \psi(x_j) \right\rangle = \sum_{i=1}^m \alpha_i \langle \psi(x_i), \psi(x_j) \rangle = \sum_{i=1}^m \alpha_i K(x_i, x_j)$$

and

$$\|w\|^2 = \left\langle \sum_{i=1}^m \alpha_i \psi(x_i), \sum_{j=1}^m \alpha_j \psi(x_j) \right\rangle = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j \langle \psi(x_i), \psi(x_j) \rangle = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(x_i, x_j)$$

Example 23. In Soft SVM (Proposition 24, Handout 6), we can have

$$\underset{\alpha}{\text{minimize}} \left(\lambda \sum_i \sum_j \alpha_i \alpha_j K(x_i, x_j) + \frac{1}{m} \sum_i \max \left(0, 1 - y_i \sum_j \alpha_j K(x_i, x_j) \right) \right)$$

for $h(x) = \text{sign} \left(\sum_j \alpha_j K(x_i, x_j) \right)$. It can be minimised via SGD. As we can just directly specify the kernel function without the need to even think about feature mapping $\psi(\cdot)$ (which is eliminated and replaced by the kernel), we can call it Kernel SVM.

Note 24. In Learning problem 18, direct access to elements $\psi(\cdot)$ in the feature space is not necessary, as equivalently one can calculate or just specify the associated kernel function (that is inner products in the feature space).

Example 25. (Polynomial Kernels) Let $x \in \mathcal{X} \subseteq \mathbb{R}^n$. Assume we want to extend the linear mapping $x \mapsto \langle w, x \rangle$ to the k degree polynomial mapping $x \mapsto h(x)$. The multivariate polynomial can be written as $h(x) = \langle w, \psi(x) \rangle$, where $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^d$ with $\psi(x)$ is a vector of elements $\psi_J(x) = \prod_{i=1}^r x_{J_i}$ for $J \in \{1, \dots, n\}^r$ and $r \leq k$. This learning problem can be equivalently be addressed with the k degree polynomial kernel

$$K(x, x') = (1 + \langle x, x' \rangle)^k$$

Solution. It is

$$\begin{aligned}
K(x, x') &= (1 + \langle x, x' \rangle)^k = \left(\sum_{j=0}^n x_j x'_j \right)^k \quad (\text{by setting } x_0 = x'_0 = 1) \\
&= \sum_{J \in \{1, \dots, n\}^k} \prod_{i=0}^n x_{J_i} x'_{J_i} = \sum_{J \in \{1, \dots, n\}^k} \left(\prod_{i=0}^n x_{J_i} \right) \left(\prod_{i=0}^n x'_{J_i} \right) \\
&= \langle \psi(x), \psi(x') \rangle
\end{aligned}$$

where $\psi(x)$ is as defined.

Example 26. (Radial basis kernel) Let the original input space be $x \in \mathcal{X} \subseteq \mathbb{R}$. Consider the Radial Basis Functions Kernel (or Gaussian kernel)

$$K(x, x') = \exp \left(-\frac{1}{2\sigma^2} \|x - x'\|_2^2 \right).$$

Show that it is a kernel indeed, by presenting it as an inner product in a feature space of infinite dimension, and state the bases of the mapping $\psi(\cdot)$.

Solution. It is

$$\begin{aligned}
K(x, x') &= \exp \left(-\frac{1}{2\sigma^2} \|x - x'\|_2^2 \right) = \exp \left(\frac{1}{\sigma^2} xx' - \frac{1}{2} x^2 - \frac{1}{2} (x')^2 \right) \\
&= \exp \left(\frac{1}{\sigma^2} xx' \right) \exp \left(-\frac{1}{2\sigma^2} x^2 \right) \exp \left(-\frac{1}{2\sigma^2} (x')^2 \right) \\
&= \sum_{k=0}^{\infty} \frac{(xx'/\sigma^2)^k}{k!} \exp \left(-\frac{1}{2\sigma^2} x^2 \right) \exp \left(-\frac{1}{2\sigma^2} (x')^2 \right) \\
&= \sum_{k=0}^{\infty} \left[\frac{x^k}{\sqrt{k!}\sigma^k} \exp \left(-\frac{1}{2\sigma^2} x^2 \right) \right] \left[\frac{(x')^k}{\sqrt{k!}\sigma^k} \exp \left(-\frac{1}{2\sigma^2} (x')^2 \right) \right]
\end{aligned}$$

hence it is $K(x, x') = \langle \psi(x), \psi(x') \rangle$ with $\psi_k(x) = \frac{x^k}{\sqrt{k!}\sigma^k} \exp \left(-\frac{1}{2\sigma^2} x^2 \right)$.

4. CONSTRUCTION OF KERNELS

Note 27. The kernel formulated as an inner product in a feature space allows us to build interesting extensions of many well-known algorithms by making use of the kernel trick and without the need to have direct access to the feature space.

Note 28. Specifying a kernel function is a way to express prior knowledge without the need to have direct access to the feature space. This is consequence of the Representation theorem 21 that kernel is the inner product of feature mappings ψ which sufficiently replaces them in the learning problem, and the fact that ψ is a way to express and utilize prior knowledge about the problem at hand.

Note 29. The question is whether the specified function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ (by the practitioner) is indeed a kernel function; i.e. if K can be written as inner product $K(x, x') = \langle \psi(x), \psi(x') \rangle$ of feature functions $\psi(x)$. Theorem 32 provides sufficient and necessary conditions to check that.

Definition 30. Gram matrix is called the $m \times m$ matrix G s.t. $[G]_{i,j} = K(x_i, x_j)$.

Definition 31. A symmetric function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is positive semi-definite if its Gram matrix G , $[G]_{i,j} = K(x_i, x_j)$, is a positive semi-definite matrix.

Theorem 32. A symmetric function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ implements an inner product in some Hilbert space is a valid kernel function in terms of if and only if it is positive semi-definite i.e. its Gram matrix G , $[G]_{i,j} = K(x_i, x_j)$, is a positive semi-definite matrix.

Proof. Assume K is a valid kernel function (i.e. it implements an inner product in some Hilbert space) $K(x, x') = \langle \psi(x), \psi(x') \rangle$; let's consider $\psi : \mathcal{X} \rightarrow \mathbb{R}^d$ for simplicity. Let G be its Gram matrix with $G = \Psi^\top \Psi$ and $\psi(x_i)$ is the i -th column of Ψ . For any $\xi \in \mathbb{R}^d - \{0\}$

$$\begin{aligned}\xi^\top G \xi &= \sum_i \sum_j \xi_i K(x_i, x_j) \xi_j = \sum_i \sum_j \xi_i \langle \psi(x_i), \psi(x_j) \rangle \xi_j = \sum_i \sum_j \langle \xi_i \psi(x_i), \psi(x_j) \xi_j \rangle \\ &= \left\langle \sum_i \xi_i \psi(x_i), \sum_j \psi(x_j) \xi_j \right\rangle = \left\| \sum_i \xi_i \psi(x_i) \right\|_2^2 \geq 0\end{aligned}$$

Assume the symmetric function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is positive semi-definite. Let $\mathbb{R}^f = \{f : \mathcal{X} \rightarrow \mathbb{R}\}$. For $x \in \mathcal{X}$ let function ψ over \mathbb{R}^f with $\psi(x) = K(\cdot, x)$. This allows to define a vector space consisting of all the linear combinations of elements of the form $K(\cdot, x)$, having an inner product

$$\left\langle \sum_i \alpha_i K(\cdot, x_i), \sum_j \beta_j K(\cdot, x_j) \right\rangle = \sum_i \sum_j \alpha_i \beta_j \underbrace{\langle K(\cdot, x_i), K(\cdot, x_j) \rangle}_{=K(x_i, x_j)}.$$

This satisfies all the properties of inner product, s.t. it is symmetric, linearity, positive definite as $K(x, x') \geq 0$. Then there is some feature vector ψ such that $K(x, x') = \langle \psi(x), \psi(x') \rangle$. \square

Claim 33. In Figure 1.1c, we could see that examples can be distinguished by some ellipse, so it was reasonable to we can define ψ as a vector with elements all the monomials up to order ; alternatively we could use a degree 2 polynomial kernel.

Proposition 34. A powerful technique for constructing new kernels is to build them out of simpler kernels as building blocks. Below are some properties. Assume $K_1 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ are valid kernels, then the following are kernels too

- (1) $K(x, x') = K_1(x, x') + K_2(x, x')$
- (2) $K(x, x') = K_1(x, x') K_2(x, x')$
- (3) $K(x, x') = K_1(x_1, x'_1) + K_2(x_2, x'_2)$, where $x = (x_1, x_2)^\top$, $x' = (x'_1, x'_2)^\top$
- (4) $K(x, x') = K_1(x_1, x'_1) K_2(x_2, x'_2)$, where $x = (x_1, x_2)^\top$, $x' = (x'_1, x'_2)^\top$
- (5) $K(x, x') = f(x) K_1(x, x') f(x')$ for any function f
- (6) $K(x, x') = K_1(f(x), f(x'))$ for any function f

Solution. We present the first two and the rest are proved similarly.

For (1). Let Gram matrix, G_j induced by kernel function K_j . For any $\xi \in \mathbb{R}^d - \{0\}$

$$\xi^\top G_3 \xi = \xi^\top (G_1 + G_2) \xi = \xi^\top G_1 \xi + \xi^\top G_2 \xi \geq 0$$

For (2). Assume that $K_j(x, x') = (\psi_j(x))^\top \psi_j(x)$. Then

$$\begin{aligned} K(x, x') &= K_1(x, x') K_2(x, x') = (\psi_1(x))^\top \psi_1(x') (\psi_2(x))^\top \psi_2(x') \\ &= (\psi_1(x))^\top \psi_2(x) (\psi_1(x'))^\top \psi_2(x') = ((\psi_1(x))^\top \psi_2(x))^\top (\psi_1(x'))^\top \psi_2(x') \end{aligned}$$

which can be represented as an inner product of feature vectors.

Note 35. The concept of a kernel formulated as an inner product in a feature space allows us to build interesting extensions of many well-known algorithms by making use of the kernel trick. One example was the Kernel SVM. Some other popular cases are Gaussian process regression, and Kernel PCA.

Draft Handout 7: Kernel methods

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce the ideas of learning machines by introducing data into high-dimensional feature spaces for accuracy gains; introduce the kernel trick, and kernel functions.

Reading list & references:

- (1) Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.
 - Ch. 6.4 Gaussian process
- (2) Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning (Vol. 1, p. 159). Cambridge, MA: MIT press.
 - Chapter 2, Regression (supplementary)

1. INTRO AND MOTIVATION

Note 1. Consider the predictive rule $h(x) = \eta(x)$. Assume there is available a set of observables $\{z_i = (x_i, y_i)\}_{i=1}^n$. We associate the learning problem with the Bayesian linear regression model (1.1)

$$\left\{ \begin{array}{l} y_i | \psi(x_i), w, \sigma^2 \stackrel{\text{ind}}{\sim} N(\eta(x_i), \sigma^2), \quad i = 1, \dots, n \\ \eta(\cdot) = (\psi(\cdot))^\top w \\ w \sim N(\mu_0, V_0) \end{array} \right. \quad \text{equiv.} \quad \left\{ \begin{array}{ll} y | \eta, \sigma^2 \sim N(\eta, I\sigma^2) & (\text{sampl. distr.}) \\ \eta = \Psi w & (\text{linear model restr.}) \\ w \sim N(\mu_0, V_0) & (\text{prior}) \end{array} \right.$$

where $[\Psi]_{i,j} = \psi_j(x_i)$.

Note 2. The marginal likelihood is

$$(1.2) \quad f(y) = N\left(y | \Psi^\top \mu_0, \Psi V_0 \Psi^\top + I\sigma^2\right)$$

where $N(y|\mu, \Sigma)$ denotes the pdf of the Normal distribution with mean μ , and covariance matrix Σ .

Note 3. The predictive distribution of a new outcome y_* at a new input x_* given the observables $\{z_i = (x_i, y_i)\}_{i=1}^n$ is

$$f(y_* | x_*, \{(x_i, y_i)\}) = N\left(\mu_*(x_*), \sigma_*^2(x_*)\right)$$

with

$$(1.3) \quad \mu_* (x_*) = \psi(x_*)^\top \mu_0 + \frac{1}{\sigma^2} \underbrace{\psi(x_*)^\top V \Psi}_{K(x_*, X)} \left(\underbrace{\Psi^\top V \Psi}_{=K(X, X)} + \sigma^2 \right)^{-1} (\Psi^\top \mu_0 - y)$$

$$(1.4) \quad \sigma_*^2 (x_*) = \underbrace{\psi(x_*)^\top V \psi(x_*)}_{=K(x_*, x_*)} - \underbrace{\psi(x_*)^\top V \Psi}_{=K(x_*, X)} \left(\underbrace{\Psi^\top V \Psi}_{=K(X, X)} + \sigma^2 \right)^{-1} \underbrace{(\psi(x_*)^\top V \Psi)^\top}_{=K(X, x_*)}$$

according to Proposition 22.

Note 4. In the prior part of (1.1), let's assume $\mu_0 = 0$ (arguably) denoting complete ignorance whether $\eta(\cdot)$ is positive or negative. Then, in (1.3) and (1.4), the feature space always enters in the form inner products. In fact we can define a kernel $K(x, x') = \langle L\psi(x), L\psi(x') \rangle = \psi(x)^\top V \psi(x')$ where L is such that $V = L^\top L$, in terms of Section 4 in Handout 7: Kernel methods. We can denote $K(x_*, X) = \psi(x_*)^\top V \Psi$, and $K(x_*, x_*) = \psi(x_*)^\top V \psi(x_*)$.

Note 5. No need to memorize the formulas in (1.2), (1.3), and 1.4. The material in Notes 1, 2, and 3 is given as a motivation for the Gaussian process regression.

2. THE GAUSSIAN PROCESS REGRESSION MODEL

Definition 6. Gaussian process (GP) is a collection of random variables $\{f(x); x \in \mathcal{X}\}$, indexed by label x , where any finite collection of those variables has a multivariate normal distribution. It is fully specified by its mean and covariance functions. It is denoted as

$$f(\cdot) \sim \text{GP}(\mu(\cdot), C(\cdot, \cdot))$$

with mean

$$\mu(x) := \text{E}(f(x)), x \in \mathcal{X}$$

and covariance function

$$C(x, x') := \text{Cov}(f(x), f(x')), x, x' \in \mathcal{X}$$

Note 7. Essentially, GP is a distribution defined over functions.

Note 8. Consider a function $\eta : \mathcal{X} \rightarrow \mathbb{R}$ with $\eta(x) = \langle \psi(x), w \rangle$ where $\psi(x)$ is a vector of known bases (feature) functions mapping from the input space to the feature space \mathcal{F} , and $w \in \mathbb{R}^d$ is an unknown vector a priori following a normal distribution $w \sim N(0, V)$, where the prior mean is set to zero denoting complete uncertainty about the sign of w 's. Then the marginal $\eta(\cdot)$ follows a Normal distribution as a linear transformation of Normal variates with mean $\text{E}(\eta(x)) = 0$ and covariance $\text{Cov}(\eta(x), \eta(x')) = \psi(x)^\top V \psi(x')$ for any $x, x' \in \mathcal{X}$. Based on the Kernel trick and Definition 6, we can equivalently specify $\eta(\cdot) \sim \text{GP}(0, C(\cdot, \cdot))$ for some kernel / covariance function $C(x, x') = \psi(x)^\top V \psi(x')$.

Note 9. We introduce the concept of Gaussian process regression in the machine learning framework below.

Note 10. Consider the predictive rule $h(x) = \eta(x)$, and assume that $\eta : \mathcal{X} \rightarrow \mathbb{R}$ with unknown image (possibly up to a set of properties, we will discuss this later) and $\mathcal{X} \subseteq \mathbb{R}^d$.

Note 11. For training purposes, assume there is available a set of observables $\{z_i = (x_i, y_i)\}_{i=1}^n$ whose sampling distribution is such that

$$(2.1) \quad \begin{aligned} y_i &= \eta(x_i) + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n \\ y_i | \eta(\cdot), \sigma^2 &\stackrel{\text{iid}}{\sim} N(\eta(x_i), \sigma^2), \quad i = 1, \dots, n \end{aligned}$$

for some unknown $\sigma^2 > 0$. This (2.1) can result by considering, a quadratic loss $\ell(h, z) = \frac{1}{\sigma^2}(h - z)^2$, and sampling distribution with pdf

$$\text{pr}(y | \{x_i, y_i\}) \propto \exp \left(- \sum_{i=1}^n \ell(h, (x_i, y_i)) \right).$$

As we $\eta(x)$ is assumed to be unknown, according to the Bayesian paradigm and the observation in Note 8, we assign a GP prior on $\eta(\cdot)$

$$\eta(\cdot) \sim \text{GP}(\mu(\cdot|\beta), C(\cdot, \cdot|\phi))$$

where μ is parametrized by unknown β (e.g. $\mu(x|\beta) = x^\top \beta$), and C is parametrized by unknown ϕ (e.g. $C(x, x'|\phi) = \exp(-\|x - x'\|_2^2 / (2\phi))$; radial/Gaussian kernel). Summing up, the Bayesian model

$$\begin{cases} y_i | \eta(\cdot) & \stackrel{\text{iid}}{\sim} N(\eta(x_i), \sigma^2), \quad i = 1, \dots, n \\ \eta(\cdot) & \sim \text{GP}(\mu(\cdot|\beta), C(\cdot, \cdot|\phi)) \end{cases}$$

where the unknown tuning parameters σ^2 , β , and $\phi > 0$ are suppressed from the conditioning.

Note 12. Consider $\eta_* = \eta(X_*)$ and $\eta_{**} = \eta(X_{**})$ where X_*, X_{**} are vectors of any length of new inputs. The joint distribution of $(\eta_*, \eta_{**}, y)^\top$ is

$$(2.2) \quad \begin{pmatrix} \eta_* \\ \eta_{**} \\ y \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C(X_*, X_*) + \sigma^2 & C^\top(X_{**}, X_*) & C^\top(X, X_*) \\ C(X_{**}, X_*) & C(X_{**}, X_{**}) + \sigma^2 & C^\top(X, X_{**}) \\ C(X, X_*) & C(X, X_{**}) & C(X, X) + \sigma^2 \end{pmatrix} \right)$$

where C is a Gram matrix such as $[C(X, X)]_{i,j} = C(x_i, x_j)$.

Note 13. Consider X_*, X_{**} as vectors of any length of new inputs. The conditional distribution of $\eta_* = \eta(X_*)$ given the training sample $\{z_i = (x_i, y_i)\}$, aka $\eta(X_*)|y$ as results from 2.2 (Proposition 22) is a normal distribution, with mean

$$\mu_*(X_*) = E(\eta_*|y) = \mu(X_*) + C(X_*, X)(C(X, X) + I\sigma^2)^{-1}(y - \mu(X))$$

at X_* and with covariance function

$$C_*(X_*, X_{**}) = \text{Cov}(\eta_*, \eta_{**}|y) = C(X_*, X)(C(X, X) + I\sigma^2)^{-1}C(X, X_{**})$$

Note 14. Given that the derivations in Note 13 is given X_* , X_{**} input vectors of any length, we can say that the predictive distribution of $\eta(\cdot)$ given the data $\{z_i = (x_i, y_i)\}$ is the Gaussian process

$$(2.3) \quad \eta(\cdot) \sim \text{GP}(\mu_*(\cdot), C_*(\cdot, \cdot))$$

with mean function and covariance function

$$\begin{aligned}\mu_*(x_*) &= \mu(x_*) + C(x_*, X)(C(X, X))^{-1}(y - \mu(X)) \\ C_*(x_*, x_{**}) &= C(x_*, X)(C(X, X))^{-1}C(X, x_{**})\end{aligned}$$

at any new points x_* , and x_{**} .

Note 15. Note that

$$\begin{aligned}\mathbb{E}(h(x_*)|y) &= \mu(x_*) + C(x_*, X)(C(X, X))^{-1}(y - \mu(X)) \\ &= \sum_{i=1}^n \alpha_i C(x_i, x_*)\end{aligned}$$

3. TRAINING

Recall that the mean and covariance functions in (2.3) depend on tunable parameters σ^2 , ϕ , and β . When the number of training examples is small, the behavior of (2.3) is sensitive to these hyperparameters.

APPENDIX A. MULTIVARIATE NORMAL DISTRIBUTION $x|\mu, \Sigma \sim N_d(\mu, \Sigma)$ ¹

Definition 16. A d -dimensional random variable $x \in \mathbb{R}^d$ is said to have a multivariate Normal (Gaussian) distribution, if for every d -dimensional fixed vector $\alpha \in \mathbb{R}^d$, the random variable $\alpha^\top x$ has a univariate Normal (Gaussian) distribution.

Definition 17. We denote the d -dimensional Normal distribution with mean μ and covariance matrix $\Sigma \geq 0$ as $N_d(\mu, \Sigma)$.

Notation 18. The d -dimensional standardized Normal distribution is $N_d(0, I)$.

Proposition 19. Let random variable $x \sim N_d(\mu, \Sigma)$, fixed vector $c \in \mathbb{R}^q$ and fixed matrix $A \in \mathbb{R}^{q \times d}$. The random vector $y = c + Ax$ has distribution $y \sim N_q(c + A\mu, A\Sigma A^\top)$.

Proposition 20. Let a d -dimensional random vector $x \sim N_{(any)}(\mu, \Sigma)$.

- (1) Let $y = Ax$ and $z = Bx$, where $A \in \mathbb{R}^{q \times d}$ and $B \in \mathbb{R}^{k \times d}$: The vectors $y = Ax$ and $z = Bx$ are independent if and only if $A\Sigma B^\top = 0$.
- (2) Let $x = (x_1, \dots, x_d)^\top$: The x_1, \dots, x_d are mutually independent if and only if the corresponding off diagonal parts of the Σ are zero.

Proposition 21. Any sub-vector of a vector with multivariate Normal distribution has a multivariate Normal distribution.

Proposition 22. [Marginalization & conditioning] Let $x \sim N_d(\mu, \Sigma)$. Consider partition such that

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix},$$

where $x_1 \in \mathbb{R}^{d_1}$, and $x_2 \in \mathbb{R}^{d_2}$ Then:

- (1) For the marginal, it is $x_1 \sim N_{d_1}(\mu_1, \Sigma_1)$.
- (2) For the conditional, if $\Sigma_1 > 0$, it is

$$x_2|x_1 \sim N_{d_2}(\mu_{2|1}, \Sigma_{2|1})$$

where

$$(A.1) \quad \mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_1^{-1}(x_1 - \mu_1) \quad \text{and} \quad \Sigma_{2|1} = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top$$

Proposition 23. The density function of the d -dimensional Normal distribution with mean μ and covariance matrix Σ , when Σ is symmetric positive definite matrix ($\Sigma > 0$), exists and it is equal to

$$(A.2) \quad f(x) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

¹More detailed material about the Multivariate Normal distribution can be found in the can be found in “Handout 2: Revision in mixture of probability distributions” of the module “Bayesian Statistics III/IV (MATH3341/4031)” Michaelmas term, 2021 available from https://github.com/georgios-stats/Bayesian_Statistics_Michaelmas_2021/blob/main/Lecture_handouts/02_Revision_in_mixture_of_probability_distributions.pdf. The material in this section is just a sub-set of the statements in the referenced handout.