Simple Termination Criteria for Stochastic Gradient Descent Algorithm

by

Sina Baghal

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Combinatorics and Optimization

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Mahdi Soltanolkotabi Assistant Professor, Departments of Electrical and Computer Engineering and Computer Science, University of Southern California
Supervisor:	Stephen Vavasis Professor, Department of Combinatorics and Optimization, University of Waterloo
Internal Members:	Joseph Cheriyan, Henry Wolkowicz Professor, Department of Combinatorics and Optimization, University of Waterloo
Internal-External Member:	Aukosh Jagannath Assistant Professor, Department of Statistics and Actuarial Science, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Stochastic gradient descent (SGD) algorithm is widely used in modern mathematical optimization. Because of its scalability and ease of implementation, SGD is usually preferred to other methods including the gradient descent algorithm in the large scale optimization. Similar to other iterative methods, SGD also needs to be employed in conjunction with a strategy to terminate the algorithm in order to prevent a phenomenon called overfitting. As overfitting is prevalent in supervised machine learning and noisy optimization problems, developing simple and practical termination criteria is therefor important. This thesis focuses on developing simple termination criteria for SGD for two fundamental problems: binary linear classification and least squares deconvolution.

In the binary linear classification problem, we introduce a new and simple termination criterion for SGD applied to binary classification using logistic regression and hinge loss with constant step-size $\alpha > 0$. Precisely, we terminate the algorithm once the margin is at least to 1. Namely,

$$\text{Terminate when } (2y_{k+1} - 1)\boldsymbol{\zeta}_{k+1}^T \boldsymbol{\theta}_k \geq 1$$

where $\boldsymbol{\theta}_k$ is the current iterate of SGD and $(\boldsymbol{\zeta}_{k+1}, y_{k+1})$ is the sampled data point at the next iteration of SGD. Notably, our proposed criterion adds no additional computational cost to the SGD algorithm. We analyze the behavior of the classifier at termination, where we sample from a normal distribution with unknown means $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \boldsymbol{R}^d$ and variances $\sigma^2 I_d$. Here $\sigma > 0$ and I_d is the $d \times d$ identity matrix. As such, we make no assumptions on the separability of the data set. When the variance is not too large, we have the following results:

1. The test will be activated for any fixed positive step-size. In particular, we establish an upper bound for the expected number of iterations before the activation occurs. This upper bound tends to a numeric constant when σ converges to zero. In fact, we show that the expected time until termination decreases linearly as the data becomes more separable (*i.e.*, as the noise $\sigma \to 0$).

2. We prove that the accuracy of the classifier at termination nearly matches the accuracy of an optimal classifier. Accuracy is the fraction of predictions that a classification model got right while an optimal classifier minimizes the probability of misclassification when the sample is drawn from the same distribution as the training data.

When the variance is large, we show that the test will be activated for a sufficiently small step-size. Finally, we empirically evaluate the performance of our termination criterion

versus a baseline competitor. We compare performances on both synthetic (Gaussian and heavy-tailed t-distribution) as well as real data sets (MNIST 51 and CIFAR-10 49). In our experiments, we observe that our test yields relatively accurate classifiers with small variation across multiple runs.

The termination criteria for SGD for the least squares deconvolution problem has not been studied in the previous literature. In this thesis, we study the SGD algorithm with a fixed step size α applied to the least square deconvolution problem [34]. We adopt the setting wherein the blurred image is contaminated with a Gaussian white noise. Under this model, we first demonstrate a novel concentration inequality which shows that for small enough step size α, the SGD path should follow the gradient flow trajectory with overwhelming probability. Inspired by numerical observation, we propose a new termination criterion for SGD for the least squares deconvolution. As a first step towards developing theoretical guarantees for our termination criterion, we provide an upper bound for the ℓ_2-error term for the iterate at termination when the gradient descent algorithm is considered. We postpone a full analysis of our termination criterion to future work.

Acknowledgements

First and foremost, I would like to thank my advisor Stephen Vavasis. Thank you for your unwavering support and your patience with me. I'm truly grateful for sharing your wisdom with me and will always appreciate the extra hours you put for reading my papers and presentations and giving me feedback and new ideas. I would like to thank my thesis committee, Henry Wolkowicz, Joseph Cheriyan, Mahdi Soltanolkotabi and Aukosh Jagannath for putting the time and effort to read my thesis and for their valuable feedback on this document. Thank you Henry for your numerous constructive comments.

I am also thankful to my friends in Waterloo for the cheerful moments we had together. Many thanks to Benjamin, Brett, Kazuhiro, Mahdi, Chris, Ali, Soroush, Advaith, Rose, Weston and of course Sir Winston. Special thanks to my friends in the US, Khashayar, Shahab (Abbas), Mahmood and Reza, for the numerous mirthful phone conversations. I am grateful for all the assistance from the members of the department administration, specifically Melissa Cambridge.

To my family

Table of Contents

List of Figures

Notation

We use the following mathematical notation in this thesis:

- d-dimensional Euclidean space is denoted by $\boldsymbol{R}^d$.
- bold-faced variable are vectors *e.g.,* $\boldsymbol{x}, \boldsymbol{u}, \cdots$. Coordinates of $\boldsymbol{x}$ are denoted by underscore notation $\boldsymbol{x}_i$.
- transpose of a vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}^T$.
- d by d identity matrix is denoted by I_d.
- diag : $\boldsymbol{R}^d \to \boldsymbol{R}^{d \times d}$ places each input vector on the diagonal.
- the inner product between $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}^d$ is denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ or $\boldsymbol{x}^T \boldsymbol{y}$.
- the i^{th} row or column of a matrix A is denoted by $A[i, :]$ and $A[:, i]$ respectively.
- $\|\boldsymbol{x}\|$, $\|\boldsymbol{x}\|_2$ or $\|\boldsymbol{x}\|_F$ denote the 2-norm or Frobenius norm of $\boldsymbol{x}$ *i.e.,* $\|\boldsymbol{x}\|_F := \sqrt{\sum_{i=1}^{d} x_i^2}$.
- $\nabla f(\boldsymbol{x})$ denotes the gradient of a function $f : \boldsymbol{R}^d \to \boldsymbol{R}$ at $\boldsymbol{x}$.
- $A \succeq 0$ means that first A is a symmetric matrix and second $\boldsymbol{x}^T A \boldsymbol{x} \geq 0$ for all $\boldsymbol{x}$.
- $N(\boldsymbol{\mu}, \Sigma)$ denotes the normal distribution with $\boldsymbol{\mu}$ and covariance matrix Σ.
- $\mathbb{E}[X]$ and $\mathrm{Var}[X]$ denotes the expected value and the variance of a random variable X. Thus, $\mathrm{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2]$.
- $\boldsymbol{\zeta} \sim \mathcal{P}$ means that the random variable $\boldsymbol{\zeta}$ follows the probability distribution $\mathcal{P}$. We also say that $\boldsymbol{\zeta}$ is sampled from the distribution $\mathcal{P}$.
- sign : $\boldsymbol{R} \to \{-1, +1\}$ is the sign function.
- $a \wedge b := \min\{a, b\}$.

Chapter 1

Introduction

This thesis studies the termination criteria for the stochastic gradient descent (SGD) algorithm arising in supervised learning and least-squares problem. The SGD algorithm is one of the most widely used iterative methods in modern optimization and machine learning. In large-scale optimization problems, SGD is usually preferred to other methods including gradient descent because of its scalability and ease of implementation [89, 10, 11, 12]. Furthermore, as a consequence of a phenomenon called overfitting any iterative algorithm must be stopped once the model has reached some desirable accuracy. The strategies based on which algorithms are halted are called termination criteria. Termination criteria are the most commonly used techniques to prevent overfitting due to their simplicity and effectiveness [38].

The main purpose of this chapter is to introduce and describe the basic concepts that we need throughout this thesis. In Section 1.1, we describe mathematical optimization. In Section 1.2, we explain the concept of termination criteria for iterative methods with a focus on gradient based algorithms such as gradient descent. In Section 1.3, we introduce the SGD algorithm. After that, in Section 1.4, we provide an explanation of supervised machine learning and the overfitting phenomenon. In Section 1.5, we discuss the least squares (LS) problem. Particularly, we will mention Tikhonov regularization and termination criteria for the LS problem in the presence of noise. Section 1.6 describes the difference between termination criteria for SGD and GD. Moreover, it provides a brief history of termination criteria for the SGD algorithm. Finally, in Section 1.7, we will outline the subsequent chapters of this thesis.

1.1 Mathematical optimization

Mathematical optimization is about finding the minima or maxima of a given function. We use the following notation:

- $\boldsymbol{x}$ is called the variables or parameters
- f is the objective or cost function
- c_i are the constraint functions, which determine specific set of equations or inequalities that the variable $\boldsymbol{x}$ must satisfy. Denote by $\mathcal{E}$ and $\mathcal{I}$ the set of indices for equality and inequality constraints, respectively.

With this notation, the optimization problem can be formulated as follows.

$$\min_{\boldsymbol{x}\in\boldsymbol{R}^d} f(\boldsymbol{x}) \quad \text{s.t.} \quad \begin{array}{ll} c_i(\boldsymbol{x}) = 0, & i \in \mathcal{E} \\ c_i(\boldsymbol{x}) = 0, & i \in \mathcal{I}. \end{array} \tag{1.1}$$

Computational algorithms used for solving (1.1) are divided into two categories: direct and iterative methods. Direct methods such as the Simplex method are not discussed in this thesis. We are primarily focusing on the iterative methods.

In iterative methods, a sequence of approximations (called iterates) are generated where each approximation is derived from the previous ones. Different iterative algorithms differ in their strategies to move from one iterate to the next one. Almost all these algorithms use the value of the objective function f, possibly its first and second derivatives and the set of constraints c_i. A good algorithm should be robust and accurate. In other words, its performance should not overly sensitive to the starting point and the errors in the data. Moreover, efficiency is always important in optimization meaning that algorithms shall not require excessive computer time or storage.

The most basic iterative method is the gradient descent (GD) algorithm (Algorithm 1) which works by iteratively moving in the opposite direction of the gradient of the function at the current iterate.

Conjugate gradient (CG) algorithm is another important iterative method which perform updates by moving along the conjugate directions. The first variant of CG methods was first proposed in the 1950s [41] as a new way for finding solutions to symmetric quadratic equations *i.e.,*

$$\min_{\boldsymbol{x}\in\boldsymbol{R}^d} f(\boldsymbol{x}) := \frac{1}{2} \|A\boldsymbol{x} - \boldsymbol{b}\|^2, \tag{1.2}$$

Algorithm 1: Gradient descent algorithm to solve $\min_{\boldsymbol{x}\in\boldsymbol{R}^d} f(\boldsymbol{x})$

Initialize: $\boldsymbol{x}_0 \in \boldsymbol{R}^d$, $\alpha > 0$
Set $k \leftarrow 0$
Repeat until a stopping criterion is satisfied
    Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)$
    $k \leftarrow k + 1$
end

Optimization problem (1.2) is called the *least squares* problem. Algorithm 2 describes the CG algorithm. In many interesting cases, CG or its variants has the opportunity to converge to the solution of (1.2) fast [65].

Algorithm 2: Conjugate gradient method for solving (1.2)

Initialize: $\boldsymbol{x}_0 \in \boldsymbol{R}^d$,
Set $\tilde{A} = A^T A$, $\boldsymbol{r}_0 \leftarrow \tilde{A}\boldsymbol{x}_0 - \boldsymbol{b}$, $\boldsymbol{p}_0 \leftarrow -\boldsymbol{r}_0$, $k \leftarrow 0$
Repeat until a stopping criterion is satisfied

$$\alpha_k \leftarrow \frac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{p}_k^T \tilde{A} \boldsymbol{p}_k}$$
$$\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$$
$$\boldsymbol{r}_{k+1} \leftarrow \boldsymbol{r}_k + \alpha_k \tilde{A} \boldsymbol{p}_k$$
$$\beta_{k+1} \leftarrow \frac{\boldsymbol{r}_{k+1}^T \boldsymbol{r}_{k+1}}{\boldsymbol{r}_k^T \boldsymbol{r}_k}$$
$$\boldsymbol{p}_{k+1} \leftarrow -\boldsymbol{r}_{k+1} + \beta_{k+1} \boldsymbol{p}_k$$
$$k \leftarrow k + 1$$

end

In contrast to GD or CG, quasi-Newton methods perform updates using the second-order information of the function f *i.e.,* the Hessian $\nabla^2 f(\boldsymbol{x})$ or its approximations. Because of this, quasi-Newton methods are called second-order methods. In this thesis, we only consider gradient based methods where the update rules at each iteration are constructed using only first order information *i.e.,* $f(\boldsymbol{x})$, $\nabla f(\boldsymbol{x})$ and etc.

Algorithms such as GD or CG are also called deterministic where the output of the model is fully determined by the parameter values and the initial conditions. In these algorithms, at each iteration the function value or its derivatives are computed to determine the next step. However, we commonly face optimization problems where the model is not

fully specified and as a result ∇f is not computable. This could be due to the fact that the function value depends on some information which will be received in the future (think about a financial portfolio optimization problem where the future interest rates are unknown). A typical way of modeling these optimization problem is by way of expressing the objective function in a form of the expectation. In doing so, we assume that the future uncertainty follows some probability distribution $\mathcal{P}$. For instance suppose that $\boldsymbol{\zeta}$ represents the uncertain parameter in our model. Assuming that $\boldsymbol{\zeta}$ follows some probability distribution $\mathcal{P}$ *i.e.,* $\boldsymbol{\zeta} \sim \mathcal{P}$, we can formulate our optimization problem in the following form:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{P}}\left[F(\boldsymbol{x}, \boldsymbol{\zeta})\right] \tag{1.3}$$

Here the objective function is a multidimensional integral and presumably it cannot be computed with a high accuracy. It is important to underline that modeling the unknown distribution $\mathcal{P}$ is quite task-specific. Nevertheless, in the absence of any prior knowledge about the distribution $\mathcal{P}$, the *normal distribution* is an appropriate default choice. This is primarily due to the fact that normal distributions exhibit the maximum amount of uncertainty once the mean and variance are fixed. Figure 1.1.

Figure 1.1: The normal curve over real numbers with mean 0 and variance 1. A random variable which follows this probability distribution is denoted by $z \sim N(0, 1)$. It is also said that z is *sampled* from the normal distribution. Normal distributions will be discussed more in Section 2.2.

As argued above, deterministic algorithms such as GD may not be a sensible/possible choice for solving (1.3) and instead stochastic algorithms are employed. The most basic

algorithm for solving (1.3) is called the stochastic gradient descent algorithm (SGD) which will be explained in the subsequent sections. We will explain SGD in two different situations wherein the distribution $\mathcal{P}$ has finite or infinite support. It is emphasized that stochastic algorithms produce solutions that optimize the *expected* performance of the model.

1.2 Termination criteria in iterative algorithms

All the iterative optimization algorithms include a stopping criterion *i.e.,* the condition for halting the algorithm. Appropriate termination criteria are of utmost importance in optimization as they save computational cost while securing solutions with high accuracy. Some termination criteria used commonly in optimization algorithms include:

- Terminate if iteration count has reached some prespecified maximum value.
- Terminate if absolute function convergence criterion is satisfied *e.g.,*

$$f(\boldsymbol{x}_k) \leq \text{ABSTOL} \cdot f(\boldsymbol{x}_0).$$

- Terminate if change in the function value in consecutive iterations is relatively small *i.e.,*

$$|f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k-1})| \leq \text{ABSFTOL} \cdot f(\boldsymbol{x}_0).$$

- Terminate if CPU time exceeds some prespecified value.

Upon a good knowledge of the problem under consideration and the employed algorithm, a specific termination criterion might be preferred to the ones listed above. For example, in unconstrained convex optimization (Section 2.1.1) since solving (1.1) is equivalent to finding $\boldsymbol{x}^*$ such that

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0},$$

the following stopping criterion is commonly used:

$$\text{Terminate when } \|\nabla f(\boldsymbol{x})\| \leq \epsilon \|\nabla f(\boldsymbol{x}_0)\|, \tag{1.4}$$

where $\epsilon > 0$ is a user-chosen parameter. Here for simplicity we assume that the optimization problem is unconstrained *i.e.,* $\mathcal{E} = \emptyset$ and $\mathcal{I} = \emptyset$. We should note in passing that the termination criterion (1.4) is also widely used in non-convex optimization problems. Also, it is worth noting that the termination criterion (1.4) is *scale-invariant*. In other words,

changing the function under consideration by $f \leftarrow \lambda f$ for some $\lambda > 0$ will not impact (1.4). It is desirable for the stopping criterion to satisfy the scale-invariance property.

The least squares problem is a special case of convex optimization. Termination criteria for (1.2), in particular for the case where the matrix A is *sparse*, have been well studied. Here sparse means that only a small fraction of entries of A are non-zero. Some few examples of this line of work include [84, 22, 85, 19, 4] where, in each of them, specific stopping criteria are suggested for halting the algorithm.

1.3 Stochastic gradient descent (SGD)

Stochastic gradient descent and its variants play a key role in modern optimization. The early history of SGD could be referenced back to the work [77]. Typical situations where SGD is used is where the objective function has a finite-sum structure such as the following.

$$\min_{\boldsymbol{x} \in \boldsymbol{R}^d} f(\boldsymbol{x}) := \frac{1}{m} \sum_{i=1}^{m} f_i(\boldsymbol{x}). \tag{1.5}$$

It should be clear that the optimization problem (1.5) is a special case of (1.3) where the support of the distribution $\mathcal{P}$ is finite and the corresponding distribution is uniform *i.e.,*

$$p(\boldsymbol{\zeta}_1) = \cdots = p(\boldsymbol{\zeta}_m) = \frac{1}{m}.$$

To obtain (1.5) from (1.3), adopt the shorthand $f_i(\boldsymbol{x}) := F(\boldsymbol{x}, \boldsymbol{\zeta}_i)$.

In certain settings, for example when m is large or the individual f_i are complicated functions, evaluating $f(\boldsymbol{x})$ or $\nabla f(\boldsymbol{x})$ can be computationally expensive. Because of this, gradient-based algorithms such as GD may not be a sensible choice for solving (1.5). On the other hand, in the presence of a large amount of uniformity in our observations, a full evaluation of $f(\boldsymbol{x})$ or $\nabla f(\boldsymbol{x})$ may not be necessary to make progress in solving (1.5). This motivates the idea that evaluating all the derivatives ∇f_i might not be necessary to perform updates and this is exactly what the SGD algorithm does. The basic idea is natural: replace the actual gradient *i.e.,* $\nabla f(\boldsymbol{x}_k)$ by an estimate thereof *i.e.,* $\nabla f_{i_k}(\boldsymbol{x}_k)$ where $i_k \sim \text{Unif}[m]$. SGD algorithm applied to (1.5) is described in Algorithm 3. Here $i \sim \text{Unif}[m]$ means that i is chosen from the set $\{1, \cdots, m\}$ uniformly at random.

It is emphasized that

$$\mathbb{E}_{i \sim \text{Unif}[m]} \left[\nabla f_i(\boldsymbol{x}) \right] = \nabla f(\boldsymbol{x}). \tag{1.6}$$

Algorithm 3: SGD algorithm for (1.5)

initialize: $\boldsymbol{x}_0 \in \boldsymbol{R}^n$, $\alpha > 0$
set $k \leftarrow 0$
repeat until a stopping criterion is satisfied
    Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla f_{i_k}(\boldsymbol{x})$ where $i_k \sim \text{Unif}[m]$
    $k \leftarrow k + 1$
end

In view of (1.6), $\nabla f_i(\boldsymbol{x})$ where $i \sim \text{Unif}[m]$ is called an unbiased stochastic gradient at the point $\boldsymbol{x}$.

Optimization problems of the form (1.5) arise in data-fitting applications where f_i corresponds to a single observation and it models the misfit of a given parameter $\boldsymbol{x}$ [38]. In modern machine learning, optimization problems in a form of (1.5) are prevalent. As a result, stochastic algorithms, in particular SGD, have been attracting a lot of attention over the last decade. Scalability for large scale models [37] and parallelizability with big training data [28] are among the most important features of the SGD algorithm.

The canonical example is the least squares problem where

$$f_i(\boldsymbol{x}) = \frac{1}{2}\left(\boldsymbol{\zeta}_i^T \boldsymbol{x} - y_i\right)^2 \qquad \text{for all } i = 1, \cdots, m.$$

Here $(\boldsymbol{\zeta}_i, y_i)$ for $i = 1, \cdots, m$ are the training data point (Section 1.4). In the event where the variable y is binary (consider $y \in \{0, 1\}$), a more appropriate model is the logistic regression model, described by the choice

$$f_i(\boldsymbol{x}) = \log\left(1 + \exp\left(-(2y_i - 1)\boldsymbol{\zeta}_i^T \boldsymbol{x}\right)\right) \qquad \text{for all } i = 1, \cdots, m. \tag{1.7}$$

It is noteworthy to mention that unlike the GD algorithm, termination criterion (1.4) cannot be used for SGD. In fact, even in the case where each f_i in (1.5) is a convex function the condition $\nabla f_{i_k}(\boldsymbol{x}_k) = \boldsymbol{0}$ does not yield optimality at $\boldsymbol{x}_k$ for the function f. Because of this, understanding termination criteria for stochastic algorithms such as SGD differ from their deterministic counterparts. We return to this issue in Section 1.6.

1.4 Supervised machine learning

The goal in supervised learning [79] is to make predictions using data. Consider the prediction problem wherein the task is to find a function f within a certain class of functions

$\mathcal{C}$ that maps from an input space $\mathcal{X}$ (for example, a set of emails) to an output space $\mathcal{Y}$ (for example, classifying those emails as spam or not-spam). We assume that we have access to a set of input-output pairs $(\boldsymbol{\zeta}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \cdots, m$ which will henceforth be called the *training dataset* such that $f^*(\boldsymbol{\zeta}_i) = y_i$ for true (unknown) $f^* \in \mathcal{C}$ for almost all $i \in \{1, \cdots, m\}$. These data are used to choose the function $\hat{f} \in \mathcal{C}$ and, assuming that there exists a good degree of uniformity between $\boldsymbol{\zeta}_i$ and y_i, then the idea is that $\hat{f}$ should provide a good prediction on subsequent pairs $(\boldsymbol{\zeta}, y) \in \mathcal{X} \times \mathcal{Y}$. The quality of the prediction that $\hat{f} : \mathcal{X} \to \mathcal{Y}$ makes on a pair $(\boldsymbol{\zeta}, y)$ is measured by way of a non-negative loss function. For example, when both $f(\boldsymbol{\zeta})$ and y are real valued, the square loss $\ell(f(\boldsymbol{\zeta}), y) := (f(\boldsymbol{\zeta}) - y)^2$ might be appropriate or, in the case where the label y takes binary values *e.g.,* $y \in \{0, 1\}$, logistic loss function (1.7) is more appropriate (Notice that the choice of the loss function ℓ is task-specific).

Upon fixing the loss function ℓ, we are interested in finding the function $\hat{f} : \mathcal{X} \to \mathcal{Y}$ in such a way that would produce a small average loss value over $(\boldsymbol{\zeta}_i, y_i)$ for $i = 1, \cdots, m$. In other words, we would like that

$$\frac{1}{m} \sum_{i=1}^{m} \ell(\hat{f}(\boldsymbol{\zeta}_i), y_i) \tag{1.8}$$

to be small. We call (1.8) the *training error* for the loss function ℓ and the training data-set $\{(\boldsymbol{\zeta}_i, y_i) : i = 1, \cdots, m\}$.

In order to ensure that $\hat{f}$ is indeed a good predictor, we assume that we have access to another collection of data points $(\tilde{\boldsymbol{\zeta}}_i, \tilde{y}_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \cdots, \tilde{m}$ which will henceforth be called the *validation dataset* and it is disjoint from the training set. We then measure how well $\hat{f}$ is making prediction on $(\tilde{\boldsymbol{\zeta}}_i, \tilde{y}_i)$. Provided that we are satisfied with the performance of $\hat{f}$ on $(\tilde{\boldsymbol{\zeta}}_i, \tilde{y}_i)$ for $i = 1, \cdots, \tilde{m}$, we could positively hope that $\hat{f}$ should predict fine on subsequent pairs $(\boldsymbol{\zeta}, y) \in \mathcal{X} \times \mathcal{Y}$. Therefore, we will evaluate the following average.

$$\frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} \ell(\hat{f}(\tilde{\boldsymbol{\zeta}}_i), \tilde{y}_i). \tag{1.9}$$

A small value of (1.9) suggests that $\hat{f}$ is indeed a good predictor. We call (1.9) the *validation error* for the loss function ℓ and the validation dataset $\{(\tilde{\boldsymbol{\zeta}}_i, \tilde{y}_i) : i = 1, \cdots, \tilde{m}\}$.

Finally, let us assume that the data pair $(\boldsymbol{\zeta}, y)$ follows a probability distribution $\mathcal{P}$ on the product space $\mathcal{X} \times \mathcal{Y}$ (Section 2.2.1, Equation 2.8). The average loss induced by $\hat{f}$ over the entire dataset is called the *generalization error*:

$$\mathbb{E}_{(\boldsymbol{\zeta}, y) \sim \mathcal{P}} \ \ell(\hat{f}(\boldsymbol{\zeta}), y) \tag{1.10}$$

Here $(\boldsymbol{\zeta}, y) \sim \mathcal{P}$ means that the random sample $(\boldsymbol{\zeta}, y)$ follows the probability distribution $\mathcal{P}$. It is worth noting that the generalization error in (1.10) can be written in the form of an integral as well *e.g.,* (1.17).

We therefore encounter two optimization problems in supervised machine learning: First, the minimization of the training error (1.8), namely

$$\min_{f \in \mathcal{C}} \frac{1}{m} \sum_{i=1}^{m} \ell(f(\boldsymbol{\zeta}_i), y_i). \tag{1.11}$$

Second, the minimization of the average loss over the entire dataset, *i.e.,*

$$\min_{f \in \mathcal{C}} \mathbb{E}_{(\boldsymbol{\zeta}, y) \sim \mathcal{P}} \ \ell(f(\boldsymbol{\zeta}), y). \tag{1.12}$$

The optimization problems (1.11) and (1.12) are called the *empirical risk minimization* and the *expected loss minimization*, respectively. There exists a trade-off between solving these two optimization problems in the following sense: Solving (1.11) to optimality to obtain $f^* \in \mathcal{C}$ does not necessary imply that f^* has a low *generalization error* meaning that $\mathbb{E}_{(\boldsymbol{\zeta}, y) \sim \mathcal{P}} \ \ell(f^*(\boldsymbol{\zeta}), y)$ is not necessarily small, especially in the presence of noise.

Let us illustrate by an example: Suppose that we are given some blue and red points as in Figure 1.2. These points are our training dataset (we can think of them as spam or not-spam emails). The task is then to find a curve such that it separates these two set of points from each other. Thus, we have training data-points $(\boldsymbol{\zeta}_i, y_i)$ where $\boldsymbol{\zeta}_i$ represents the dots in Figure 1.2 and $y_i \in \{\text{blue, red}\}$. The class of functions $\mathcal{C}$ is considered to be set of all functions that are 1 on one side of a curve and 0 on the other side. For a predictor f, the prediction $f(\boldsymbol{\zeta})$ is defined in the most obvious way. Now in order to express our task in terms of the optimization problems (1.11) and (1.12), it remains to define the loss function. Let $\ell_{0,1}$ be the 0-1 loss function, namely,

$$\ell_{0,1}(f(\boldsymbol{\zeta}), y) = \begin{cases} 0, & f(\boldsymbol{\zeta}) = y \\ 1, & f(\boldsymbol{\zeta}) \neq y. \end{cases}$$

With all these notation and definitions, we are led to believe that the black curve (denote it by f_1) must have a lower generalization error that the green curve (denote it by f_2). Nonetheless, clearly f_2 has a lower training error than f_1. As it has been illustrated in Figure 1.2, the reason that f_1 yields lower generalization error is due to the fact that f_2 is *fitting* the *noisy* training data points. By noise, we mean the data points that are not representative of the true properties of data. This phenomenon is called *overfitting*.

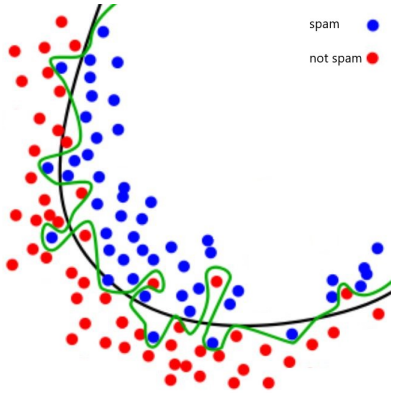

Figure 1.2: A cartoon depiction of the concept of overfitting in supervised machine learning.

The methods which are used to prevent overfitting without reducing the generalization error nor the training error are called *regularization*. In the subsequent sections, we will discuss more about regularization. The cartoon in Figure 1.2 is also an example of a binary classification problem which we will return to in Section 1.4.3.

The class of functions $\mathcal{C}$ is generally parametrized by a vector $\boldsymbol{\theta} \in \Theta$ where Θ is a subset of some fixed Euclidean space. We provide an example below from the neural network literature which are not further pursued in this thesis. Here the class of functions $\mathcal{C}$ used for training has a very specific structure.

Example 1. (Neural networks) A neural network is represented by $\boldsymbol{\theta} \in \Theta := \boldsymbol{R}^{N_1 \times N_0} \times \cdots \times \boldsymbol{R}^{N_L \times N_{L-1}}$ where $N_0 = d, N_L = 1$. Here, for an input $\boldsymbol{\zeta} \in \boldsymbol{R}^d$, the output of the network for $\boldsymbol{\theta} = (W_L, \cdots, W_1)$ is equal to

$$f_{\boldsymbol{\theta}}(\boldsymbol{\zeta}) = W_L \circ \sigma \circ W_{L-1} \circ \sigma \circ \cdots \circ W_2 \circ \sigma \circ W_1 \boldsymbol{\zeta}, \tag{1.13}$$

where σ is called the activation function which acts entry-wise on its vector input. Common activation functions are ReLU *i.e.* $\sigma(x) := \max(x, 0)$, and sigmoid *i.e.* $\sigma(x) := \frac{1}{1+e^x}$. Given samples $\{(\boldsymbol{\zeta}_i, y_i)\}_{i=1}^m$, the learning procedure involves the following empirical risk minimization (1.11) where the square loss is used

$$\ell(f_{\boldsymbol{\theta}}(\boldsymbol{\zeta}), y) = \left(y - W_L \circ \sigma \cdots \circ \sigma \circ W_1 \boldsymbol{\zeta}\right)^2. \tag{1.14}$$

Neural networks forms the basis of remarkable advances in many areas of machine learning [38].

1.4.1 SGD for general distribution

In Section 1.3, we discussed the SGD algorithm for the case where the objective function is written in a finite-sum form. However, consider the expected loss minimization problem (1.12) where the sum is taken over a general probability distribution. Can we use SGD to solve the problems of the form in (1.12)? For clarity, in the rest of this chapter, we use the parametrization $\mathcal{C} \equiv \Theta$ to represent the elements of $\mathcal{C}$ whenever it is notionally more convenient. Also, we denote

$$\ell_{\boldsymbol{\theta}}(\boldsymbol{\zeta}, y) := \ell(f(\boldsymbol{\zeta}), y),$$

where $\boldsymbol{\theta} \in \Theta$ corresponds to f. With this notation, the minimization problem (1.12) is rewritten as follows.

$$\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}} \ \ell_{\boldsymbol{\theta}}(\boldsymbol{\zeta}, y). \tag{1.15}$$

Algorithm 4 describes the SGD algorithm applied to (1.15) where at the iteration k, we sample $(\boldsymbol{\zeta}_k, y_k)$ from the distribution $\mathcal{P}$. In this setting, the training set as in Section 1.3 is replaced by an oracle that generates instances $(\boldsymbol{\zeta}_i, y_i)$ on demand for $i = 1, 2, \cdots$. Notice that this new setting encompasses the previous one by letting $\mathcal{P}$ to be the uniform distribution over the finite set

$$\{(\boldsymbol{\zeta}_i, y_i) : i = 1, \cdots, m\}.$$

It is also assumed that the class of functions $\mathcal{C}$ has been parametrized by $\boldsymbol{\theta} \in \Theta$ in such a way that the update formula

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} \ell\left(\boldsymbol{\theta}_k(\boldsymbol{\zeta}_k), y_k\right) \qquad \text{where } (\boldsymbol{\zeta}_k, y_k) \sim \mathcal{P}$$

makes sense. For example, in the binary classification problem (Section 1.4.3), each $f \in \mathcal{C}$ can be represented as follows

$$f(\boldsymbol{\zeta}) = \boldsymbol{h}^T \boldsymbol{\zeta} + b,$$

for some $(\boldsymbol{h}, b) \in \Theta := \boldsymbol{R}^d \times \boldsymbol{R}$. Thus, $\mathcal{C} \equiv \Theta = \boldsymbol{R}^d \times \boldsymbol{R}$.

Algorithm 4 converges to a neighborhood of the minimizer [70] whose size is controlled by the following variance parameter

$$\mathbb{E}\left[\left\|\nabla_{\boldsymbol{\theta}_k} \ell_{\boldsymbol{\theta}_k}\left(\boldsymbol{\zeta}_k, y_k\right) - \nabla_{\boldsymbol{\theta}_k} \mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}} \ \ell_{\boldsymbol{\theta}_k}(\boldsymbol{\zeta}, y)\right\|^2\right] \leq \tau^2, \tag{1.16}$$

for some $\tau > 0$. The bound in (1.16) holds by assumption (Section 2.1.2). In words, we assume that the stochastic gradients $\nabla_{\boldsymbol{\theta}_k} \ell_{\boldsymbol{\theta}_k}\left(\boldsymbol{\zeta}_k, y_k\right)$ and the full gradient $\nabla_{\boldsymbol{\theta}_k} \mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}} \ \ell_{\boldsymbol{\theta}_k}(\boldsymbol{\zeta}, y)$ are close to each other in the sense of (1.16). It is emphasized that after arriving at the aforementioned neighborhood, the SGD iterates start to oscillate. We provide an illustrative example next.

Algorithm 4: SGD algorithm for (1.15)

initialize: $\boldsymbol{x}_0 \in \boldsymbol{R}^n$, $\alpha > 0$
set $k \leftarrow 0$
repeat until a stopping criterion is satisfied
    Update $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} \ell_{\boldsymbol{\theta}_k} (\boldsymbol{\zeta}_k, y_k)$ where $(\boldsymbol{\zeta}_k, y_k) \sim \mathcal{P}$
    $k \leftarrow k + 1$
end

Example 2. Suppose that $\Theta = \boldsymbol{R}$ and let

$$\ell_\theta(\zeta, y) := -\theta y \zeta + \log\left(1 + \exp(\theta\zeta)\right),$$

where $\theta \in \boldsymbol{R}$ and $(\zeta, y) \sim \mathcal{P}$. In other words, we let

$$C \equiv \left\{(\zeta, y) \mapsto -\theta y \zeta + \log\left(1 + \exp(\theta\zeta)\right) : \theta \in \boldsymbol{R}\right\}.$$

We further assume that $\mathcal{P}$ follows a Gaussian mixture model:

$$\mathbb{P}(y = 1) = \mathbb{P}(y = 0) = \frac{1}{2}.$$

Once y is selected, then the marginal distribution of ζ is as follows.

$$\zeta \sim N(1 - 2y, 1).$$

This means that

$$\mathbb{P}(\zeta \le t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t - 1 + 2y)^2}{2}\right),$$

i.e., ζ is a real-valued random variable which is governed by a Gaussian distribution (Section 2.2.2). We thus have

$$\begin{aligned}\mathbb{E}_{(\zeta,y)\sim\mathcal{P}}\ \ell_\theta(\zeta, y) = &\frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{+\infty} \log\left(1 + \exp(\theta\zeta)\right) \cdot \exp\left(-\frac{(\zeta - 1)^2}{2}\right) d\zeta \\ &+ \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{+\infty} \log\left(1 + \exp(-\theta\zeta)\right) \cdot \exp\left(-\frac{(\zeta + 1)^2}{2}\right) d\zeta.\end{aligned} \tag{1.17}$$

Algorithm 4 used for solving (1.12) performs the following update rule:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \frac{\alpha(2y_k - 1)\zeta_k}{1 + \exp(\boldsymbol{\theta}_k(2y_k - 1)\zeta_k)}.$$

Here $(\zeta_1, y_1), (\zeta_2, y_2), \cdots$ is a sequence of random variables drawn from the distribution $\mathcal{P}$. We return to this setting in Chapter 3.

1.4.2 Regularization in supervised machine learning

Avoiding overfitting is a major aspect of training in supervised machine learning. Overfitting occurs when the training error is small and generalization error is large *i.e.,* the *generalization gap* is large. Consider the training and generalization error minimization in (1.8) and (1.12) respectively. Suppose that some iterative optimization *e.g.,* SGD has been used to train the desired model and denote $f_1, f_2, \cdots$ the corresponding iterates (f_k corresponds to $\boldsymbol{\theta}_k$ in Algorithm 4). Denote the training and generalization error at iteration k as follows.

$$\text{Training-error}_k := \frac{1}{m}\sum_{i=1}^{m} \ell(f_k(\boldsymbol{\zeta}_i), y_i) \text{ and } \text{Gen-error}_k := \mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}}\ell(f_k(\boldsymbol{\zeta}), y).$$

In view of the fact that increasing k, the complexity of model f_k increases, the iteration count k is also called the complexity of the model k. A typical relationship between $\{\text{Training-error}_k\}_{k=0}^{+\infty}$ and $\{\text{Gen-error}_k\}_{k=0}^{+\infty}$ is illustrated in Figure 1.3. Ideally, we need to halt our iterative algorithm once the generalization gap is small. This strategy is known as *early stopping* or *implicit regularization*.

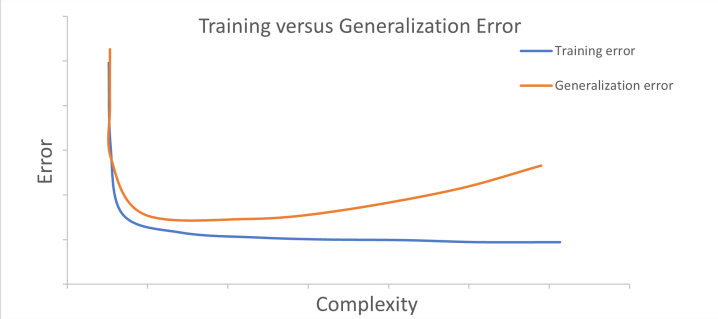

Figure 1.3: Complexity versus error.

A fundamental difference between the termination criteria we discussed in Section 1.2 and the ones in supervised machine learning is that the later are meant to produce low generalization error rather than capturing convergence. We say an algorithm exhibits a implicit regularization behaviour when for some termination criterion the generated model has low generalization error as in Figure 1.3.

Generally, regularization is any modification we make to our learning algorithm in order to reduce its generalization gap but not its training error [38]. Beside to early stopping, the most common form of regularization is by adding a penalty term to the objective function.

By doing so, we give our algorithm a preference for one solution over another. This can be formulated as follows: for some non-negative weight function $J : \mathcal{C} \to \boldsymbol{R}$, we replace the training error minimization (1.8) with the following regularized version.

$$\min_{f \in \mathcal{C}} \frac{1}{m} \sum_{i=1}^{m} \ell(f(\boldsymbol{\zeta}_i), y_i) + \lambda J(f). \tag{1.18}$$

The weight function is designed in such a way that for undesirable functions $f \in \mathcal{C}$, $J(f)$ is relatively large. Also, the parameter $\lambda > 0$ controls the intensity of the regularization term. It is emphasized that the choice of function J is made based on the prior knowledge we have about the problem under consideration. For a proper choice of λ, the orange curve in Figure 1.3 becomes more flat towards the end and hence even a very small training error should result in a low generalization error.

Regularization of the form (1.18) has a long history in optimization. In the subsequent sections, as an example, we will discuss the Tikhonov regularization for solving noisy least squares problem.

1.4.3 Binary linear classification

In a binary classification task, the goal is to specify which of the two categories some input belongs to. With the notation from Section 1.4, in a binary classification problem, we have that $|\mathcal{Y}| = 2$ (for simplicity, denote $\mathcal{Y} = \{0, 1\}$). The set of inputs *i.e.,* $\mathcal{X}$ is also partitioned as

$$\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1. \tag{1.19}$$

By definition $\mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset$. Then for any function $f : \mathcal{X} \to \mathcal{Y}$, the number of misclassifications is defined by

$$\text{error}(f) := \left| \{\boldsymbol{\zeta} \in \mathcal{X}_0 : f(\boldsymbol{\zeta}) = 1\} \cup \{\boldsymbol{\zeta} \in \mathcal{X}_1 : f(\boldsymbol{\zeta}) = 0\} \right|. \tag{1.20}$$

The goal is now to find f such that $\text{error}(f)$ is small. In supervised binary classification, we are provided with m training data pairs

$$(\boldsymbol{\zeta}_1, y_1), \cdots, (\boldsymbol{\zeta}_m, y_m), \tag{1.21}$$

where y_i denotes the correct label for $\boldsymbol{\zeta}_i$. Using these m pairs, we would like to construct a model such that for a new pair $(\boldsymbol{\zeta}, y)$, it could predict y after observing $\boldsymbol{\zeta}$. In order to made this task possible from a mathematical perspective, we need to assume that the set

of inputs $\mathcal{X}$ is structured in some particular way. To this end, we need to visualize the set of inputs using some mathematical object. This is commonly done by way of embedding $\mathcal{X}$ into some Euclidean space $\boldsymbol{R}^d$. In other words, we assume that

$$\mathcal{X} \subseteq \boldsymbol{R}^d. \tag{1.22}$$

In a binary *linear* classification task, we make one further fundamental assumption. We suppose that the embedding (1.22) satisfies the following property:

$$\mathcal{X}_0 \text{ and } \mathcal{X}_1 \text{ are almost separable by a hyperplane.} \tag{1.23}$$

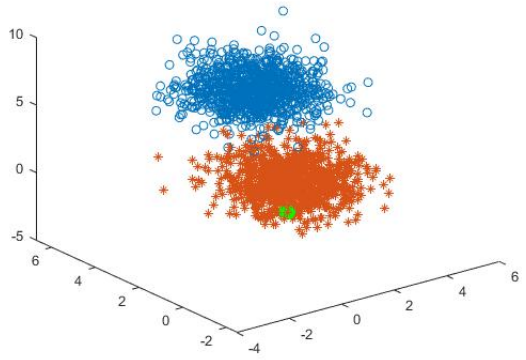

Figure 1.4: In a binary linear classification problem, the task of the learner is to find a hyperplane (linear classifier) which separates two group of points while minimizing the number of misclassifications.

Here by the word almost, we mean that it might be the case where no hyperplane would separate these two set of nodes $\mathcal{X}_0$ and $\mathcal{X}_1$ completely, but there exists some hyperplane such that the number of misclassified nodes is negligible. Assumptions (1.22) and (1.23) are illustrated in Figure 1.4. Notice that there exists a hyperplane which almost separates blue and orange nodes from each other. Define the 0-1 error loss function as follows:

$$\ell_{0,1}(z, y) = \begin{cases} 0, & \operatorname{sign}(z) = 2y - 1 \\ 1, & \operatorname{sign}(z) \neq 2y - 1. \end{cases}$$

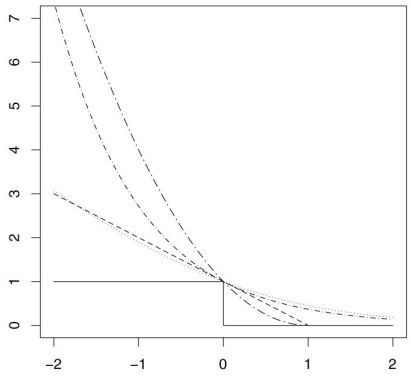

Figure 1.5: Common surrogate loss functions: exponential, hinge, logistic and truncated quadratic [5]. Here, the staircase non-convex function is the $0-1$ loss function which is equal to 0, if the model predicts correct, 1 otherwise.

With this notation, the minimization of error(f) (1.20) can be written in the following way:

$$\min_{(\boldsymbol{h},b)\in\boldsymbol{R}^d\times\boldsymbol{R}} \frac{1}{m}\sum_{i=1}^{m} \ell_{0,1}(f_{\boldsymbol{h},b}\left(\boldsymbol{\zeta}_i\right), y_i), \tag{1.24}$$

where $f_{\boldsymbol{h},b}$ is defined by

$$f_{\boldsymbol{h},b}(\boldsymbol{\zeta}) := \boldsymbol{h}^T\boldsymbol{\zeta} + b.$$

The optimization problem (1.24) is the empirical risk minimization defined in (1.8) which is often intractable (exponential in d) [53]. As a result, a surrogate loss function is used instead to replace (1.24) with a convex relaxation (Chapter 2). Some common loss functions used for the task of binary classification are plotted in Figure 1.5. It is well known that the model obtained using these surrogate loss functions provides a non-trivial upper bound on the excess risk (error associated with the 0-1 loss function) under the weakest possible condition [5]. Note that convexity of the surrogate loss functions make the solving algorithms computationally efficient.

1.5 Linear least squares problems

Given an $m \times n$ matrix A and an $n \times 1$ vector $\boldsymbol{b}$, the least squares problem is to find a vector $\boldsymbol{x}$ such that it minimizes $\|A\boldsymbol{x} - \boldsymbol{b}\|$. Therefore, we are interested in the following

optimization problem:

$$\min f(\boldsymbol{x}) := \frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|^2. \tag{1.25}$$

The least squares problem appear in many applications. A few examples include digital image restoration [34], statistical modeling [36] and curve fitting [35]. Below, we briefly discuss three standard ways for solving (1.25): the normal equations, the QR decomposition and the singular value decomposition (SVD).

It can be checked easily that f is a convex function and hence any solution $\boldsymbol{x}_{\text{LS}}$ to (1.25) must satisfy the following equation.

$$A^T A\boldsymbol{x}_{\text{LS}} = A^T\boldsymbol{b}. \tag{1.26}$$

Equation (1.26) is called the *normal equation*. We will only concentrate on the overdetermined case where $m > n$ and furthermore, we suppose that A has full column rank. Under these assumption, by (1.26), we have that

$$\boldsymbol{x}_{\text{LS}} = \left(A^T A\right)^{-1} A^T\boldsymbol{b}. \tag{1.27}$$

Since $A^T A$ is positive definite, we can use the Cholesky decomposition to obtain $\boldsymbol{x}_{\text{LS}}$ from (1.27). We next state the QR decomposition and SVD.

Fact 1. There exists a matrix Q *i.e.,* $Q^T Q = I_n$, and a unique upper triangular matrix R with non-negative diagonal entries such that $A = QR$.

Fact 2. We can write $A = U\Sigma V^T$ where $U \in \boldsymbol{R}^{m\times n}$ and $V \in \boldsymbol{R}^{n\times n}$ such that $U^T U = V^T V = I_n$, and $\Sigma = \text{diag}\,(\sigma_1, \cdots, \sigma_n)$ with $\sigma_i \geq 0$. The columns of U and V are called left and right singular vectors respectively. σ_i are called the singular values of A.

In the case where A is full column rank, it holds that R is non-singular and $\sigma_i > 0$ in the QR decomposition and SVD respectively. It is also readily verified that

$$A^\dagger := \left(A^T A\right)^{-1} A^T = R^{-1}Q = V\Sigma^{-1}U^T, \tag{1.28}$$

where $A^\dagger$ is called the *Moore-Penrose* pseudoinverse of A. It is emphasized that the QR decomposition and SVD methods are particularly helpful in the rank-deficient or ill-condition problems *i.e.,* some of singular values are either zero or very small respectively. Such problems arise in many applications and regularization methods are necessary to mitigate the bad effects of very small singular values. In the next subsection, we will discuss some of the standard regularization methods used for solving the ill-conditioned least squares problems.

1.5.1 Regularization techniques for LLS

Small singular values cause troubles when our task is solving (1.25). To illustrate, let us assume that the singular values are ordered as follows.

$$\sigma_1 \geq \cdots \geq \sigma_n \geq 0.$$

We clearly have that

$$\|\boldsymbol{x}_{\mathrm{LS}}\|_2 \geq \left| \frac{U[:,n]^T \boldsymbol{b}}{\sigma_n} \right|.$$

In particular, changing $\boldsymbol{b}$ to $\boldsymbol{b} + \epsilon U[:,n]$ change $\boldsymbol{x}_{\mathrm{LS}}$ to $\boldsymbol{x}_{\mathrm{LS}} + \tilde{\boldsymbol{x}_{\mathrm{LS}}}$ where

$$\|\tilde{\boldsymbol{x}_{\mathrm{LS}}}\|_2 \geq \frac{\epsilon}{\sigma_n}.$$

Combining pieces, we conclude that the solution to (1.25) is potentially very large and also very sensitive to error in the vector $\boldsymbol{b}$. In consequence, we say that the least squares problem (1.25) is ill-conditioned whenever $\sigma_n \approx 0$. This issue will be intensified once the vector $\boldsymbol{b}$ is noisy. To explain, let us write

$$\boldsymbol{b} = A\boldsymbol{x}^* + \boldsymbol{\zeta}, \tag{1.29}$$

where $\boldsymbol{\zeta}$ denotes the noise in our model and $\boldsymbol{x}^*$ denotes the sought-after unknown vector (when $\boldsymbol{\zeta} = 0$ in (1.29), $A^\dagger \boldsymbol{b} = \boldsymbol{x}^*$). By (1.27) and $\boldsymbol{b}$ as in (1.29), the solution to (1.25) can be computed as follows.

$$\boldsymbol{x}_{\mathrm{LS}} := \boldsymbol{x}^* + A^\dagger \boldsymbol{\zeta}$$

By (1.28), we have that

$$A^\dagger \boldsymbol{\zeta} = \sum_{i=1}^{n} \frac{U[:,i]^T \boldsymbol{\zeta}}{\sigma_i} V[:,i].$$

Similarly, we have that

$$\left| \frac{U[:,n]^T \boldsymbol{\zeta}}{\sigma_n} \right| \gg 0,$$

assuming that

$$|U[:,n]^T \boldsymbol{\zeta}| \not\approx 0. \tag{1.30}$$

The noise vector $\boldsymbol{\zeta}$ can be formulated through some mathematical modeling. For example, in Chapter 4, $\boldsymbol{\zeta}$ is modeled as an isotropic random Gaussian vector. Under this assumption, the condition in (1.30) always holds. In fact, we will see that the expected value of $|U[:$

$, n]^T \boldsymbol{\zeta}|^2$ is equal to the variance in the modeled noise. Notice that across many applications, it is not only the last singular value that might happen to be near zero as, generally, in ill-conditioned problems, for some $s > 0$ it holds that

$$\sigma_{n-s+1} \approx 0, \cdots, \sigma_n \approx 0. \tag{1.31}$$

We summarize our discussion below:

Fact 3. When A is ill-conditioned, the solution $\boldsymbol{x}_{\text{LS}} = A^\dagger \boldsymbol{b}$ is dominated by the contributions from rounding and data errors. In view of this, $A^\dagger \boldsymbol{b}$ is called the *naive solution* to (1.25).

To damp these contributions, the regularization methods are applied to the least squares problem (1.25). Notice that, in many places, it is not only the smallest singular value which is approximately zero, but it even holds that

$$U[:, n - s' + 1]^T \boldsymbol{x}^* \approx 0, \cdots, U[:, n]^T \boldsymbol{x}^* \approx 0, \tag{1.32}$$

where $s' \geq s$ where s is defined in (1.31). One such example is in the image deblurring problems (Chapter 4). In view of (1.31) and (1.32), the following truncated sum might be considered as the solution instead of $\boldsymbol{x}_{\text{LS}}$

$$\mathcal{R}_k(\boldsymbol{x}_{\text{LS}}) = \sum_{i=1}^{k} \frac{U[:, i]^T \boldsymbol{b}}{\sigma_i} V[:, i]. \tag{1.33}$$

For a properly chosen k, (1.33) produces a good approximation of $\boldsymbol{x}^*$. This method is called the truncated SVD (TSVD) method [34]. TSVD is a specific example of a broader class of methods that are called *spectral filtering* methods, which have the following form

$$\boldsymbol{x}^*_{\text{filtered}} := \sum_{i=1}^{n} f_i \frac{U[:, i]^T \boldsymbol{b}}{\sigma_i} V[:, i],$$

where f_i are called the filter factors. The main idea of spectral filtering is to choose f_i such that for large singular values $f_i \approx 1$ and $f_i \approx 0$ for small singular values. To this end, one common approach is to let

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda}, \tag{1.34}$$

where $\lambda > 0$ is a parameter which needs to be tuned. It can be easily verified that for f_i in (1.34), the filtered solution $\boldsymbol{x}^*_{\text{filtered}}$, denote it by $\boldsymbol{x}_\lambda$, is the solution to the following optimization problem.

$$\min f_\lambda(\boldsymbol{x}) := \frac{1}{2} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda \|\boldsymbol{x}\|^2. \tag{1.35}$$

The optimization problem (1.35) is called the Tikhonov regularization problem [76, 71]. In order to approximate the optimal regularized solution $\boldsymbol{x}_{\lambda_{\text{opt}}}$ *i.e.,*

$$\boldsymbol{x}_{\lambda_{\text{opt}}} := \underset{\lambda>0}{\operatorname{argmin}} \|\boldsymbol{x}_\lambda - \boldsymbol{x}^*\|,$$

we need to estimate the optimal value λ_{opt}. Trial and error schemes and the L-curve criterion [31] are among the common strategies used in practice to tune λ. See also [26, 29]. It is emphasized that the task of tuning λ always requires the computation of the regularized solution $\boldsymbol{x}_\lambda$ for some few different values of λ. Finally, it is worth noting that specialized methods have been developed for solving (1.35) where the matrix A is sparse. These instances are called sparse least squares problems and they appear in many places such as image deblurring. Some few examples of this line of work include [84, 22, 85].

The implicit regularization effect of early stopping is another form of regularization for the least squares problem. To illustrate, we provide an example of an image deblurring problem. As it will be detailed in Chapter 4, a digital image deblurring problem can be formulated in a form of a least squares problem (1.25) where $\boldsymbol{b}$ is defined in (1.29). The unknown $\boldsymbol{x}^*$ represents the deblurred (also called sharp) image, the observed vector $\boldsymbol{b}$ represents the blurred noisy image, and A is called the blurring matrix. We apply the gradient descent algorithm on (1.25) (Algorithm 5).

Algorithm 5: GD algorithm to minimize $\frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|^2$

initialize: $\boldsymbol{x}_0 = \boldsymbol{0} \in \boldsymbol{R}^n$, $\alpha > 0$
for $k = 0, 1, \cdots$
    Update $\boldsymbol{x}_{k+1}^{\text{GD}} = \boldsymbol{x}_k^{\text{GD}} - \alpha A^T \left(A\boldsymbol{x}_k^{\text{GD}} - \boldsymbol{b}\right)$
    $k \leftarrow k + 1$
end

It is worth noting that in this setting each $\boldsymbol{x}_k^{\text{GD}}$ generated by Algorithm 5 represents an image. In particular, $\boldsymbol{x}_0 = \boldsymbol{0}$ represents the pitch black image. See Figure 1.6 for an example. Let us define

$$\text{Error of } \boldsymbol{x}_k^{\text{GD}} := \frac{\|\boldsymbol{x}_k^{\text{GD}} - \boldsymbol{x}^*\|_2}{\|\boldsymbol{x}^*\|_2}.$$

From Figure 1.6, we observe that the error of $\boldsymbol{x}_k^{\text{GD}}$ decays at first (and becomes close to a regularized solution) and then it keeps ascending. In view of this, there exists an optimal termination criterion T_{opt} such that halting the algorithm at the iteration T_{opt} yields a desirable accuracy *i.e.,*

$$\text{Error of } \boldsymbol{x}_{T_{\text{opt}}}^{\text{GD}} \text{ is small.} \tag{1.36}$$

Figure 1.6: Image deblurring is the process of removing blurring artifact from images (Chapter 4). The corresponding iterates for $k = 0, 50, 100, 150, 300$ of the GD algorithm are pictured. Notice that the iterate 50 exhibits a desirable accuracy.

Designing an explicit termination criterion T such that $|T - T_{\text{opt}}|$ is not too large is therefore important. When (1.36) holds, we say that Algorithm 5 exhibits an implicit regularization effect, also known as the *semi-convergence behaviour* [59, p. 89]. The semi-convergence behaviour of iterative algorithms is well studied *e.g.,* [20, 6, 33, 40, 85, 73].

For a general iterative algorithm applied to (1.25), the implicit regularization effect, the error term and the termination criterion T are defined similarly as above. Notice that these type of termination criteria are different from what we discussed in Section 1.2, in particular the termination criteria used in the sparse least squares literature [84, 22, 85, 19, 4]. In fact, in these line of works, termination criteria are designed to capture convergence to the minimizer of the iterative algorithm being used rather that exploiting their implicit regularization effects. In Chapter 4, we will study the implicit regularization effect of the *SGD algorithm* applied to the least squares problem (1.25) arising in image deblurring. Notice that the objective function in (1.25) can be written in a form of the finite-sum consisting of m terms to which the SGD algorithm from Section 1.3 can be applied.

1.6 Termination criteria for SGD

This thesis deals with termination criteria for the SGD algorithm. Recall that in Section 1.4.2, we introduced the concept of early stopping regularization. Unlike regularization via penalization, the early stopping approach does not change the training procedure and also eliminates the required computations needed for the regularization term. In view of these pleasant statistical and computational benefits, developing useful termination criteria for iterative algorithms is therefore important.

While termination criteria for deterministic algorithms such as GD have been widely studied *e.g.,* [22, 68], in the case of stochastic algorithms such as SGD they are not well

understood yet. It is important to note that why a good termination criterion for GD does not necessarily yield a proper termination criterion for SGD. Let us illustrate by an example: In Example 2, consider the following test:

$$\text{Terminate when } \|\nabla_{\theta_k} \ell_{\theta_k}(\zeta_k, y_k)\| \leq \epsilon \left\|\nabla_{\theta_0} \mathbb{E}_{(\zeta,y)\sim\mathcal{P}} \ell_{\theta_0}(\zeta, y)\right\|. \tag{1.37}$$

Further, assume that at some iteration k, we observe that

$$\zeta_k \gg 0 \text{ and } y_k = 1. \tag{1.38}$$

Notice that

$$\|\nabla_{\theta_k} \ell_{\theta_k}(\zeta_k, y_k)\| = \left\| \frac{\zeta_k}{1 + \exp(\theta_k \zeta_k)} \right\|.$$

Hence, by (1.38), we expect (1.37) is satisfied provided that $\theta_k > 0$. In view of this, the unlikely event (1.38) might cause termination even though the iterate θ_k might not have any important statistical feature such as closeness to the minimizer or exhibiting a low error as a classifier (1.20).

Recall that SGD with a constant step-size converges to a neighborhood of the minimizer, whose size depends on the variance parameter defined in (1.16), and then starts to oscillate. Termination criteria for the SGD algorithm to diagnose convergence to the minimizer is therefore important. For SGD with a constant step-size such as Algorithms 3 and 4, in [15, 69] the authors have developed explicit and cheaply computable termination criteria to halt the iteration updates once the algorithm has arrived at a small neighborhood of the minimizer.

The SGD algorithm with shrinking step-sizes does not exhibit the oscillatory behaviour around the minimizer similar to the case where the step-size is fixed. It is well known that, with a proper policy of shrinking step-sizes, SGD is guaranteed to converge to the minimizer *e.g.,* [62] and therefore some of the termination criteria from Section 1.2 could be used to halt the algorithm once the convergence has occurred.

As argued in Sections 1.4.2 and 1.5.1, in the context of supervised machine learning or noisy least squares problem, termination criteria for iterative algorithms to detect convergence to the minimizer are not useful. In the context of supervised learning, the earliest comprehensive numerical testing of a stopping criterion for SGD was introduced in [72]. Their stopping criterion, which we call it the *small validation set* (SVS) test, periodically checks the accuracy of the iterate on the validation dataset. Every time the error on the validation set improves, a copy of the model parameters is stored. The algorithm is halted if there has been no improvement in the error on the validation set. Theoretical guarantees for SVS are established in *e.g.,* [52, 87].

1.7 Outline of the thesis

In the next chapter, we will provide some background material which we will need throughout the thesis. We arrive at our first main result in Chapter 3 where we propose a new, simple, and computationally inexpensive termination test for SGD applied to binary linear classification on the logistic and hinge loss functions. Our theoretical results support the effectiveness of our stopping criterion when the data is Gaussian distributed. We show that our test terminates in a finite number of iterations and when the noise in the data is not too large, the expected classifier at termination nearly minimizes the probability of misclassification. Next, in Chapter 4, we consider the SGD algorithm for the least squares deconvolution problem. We prove a new concentration inequality to demonstrate that the SGD algorithm shall follow the gradient flow trajectory with high probability. Based on numerical observations, we propose a computationally inexpensive termination criterion for the SGD algorithm. As a first step towards developing a theoretical understanding for our test, we provide a bound for the ℓ_2-error term for the iterate at termination for the GD algorithm. Finally in Chapter 5, we summarize the key results of the thesis and list some of the unanswered interesting questions which this work presents for the future inquiry.

Chapter 2

Preliminaries

In this chapter, we establish the basic notation and record some preliminary results that we will use throughout the thesis. We should emphasize that none of the material in this chapter is new.

2.1 Optimization

This section is devoted to some mathematical optimization background. Section 2.1.1 contains some basic definitions and examples from convex analysis. In Section 2.1.2, we discuss the convergence analysis of the SGD algorithm.

2.1.1 Convex analysis

The concept of convexity is fundamental in optimization. Convex problems are easier to solve both in theory and practice and they are present across many type of application [13, 78, 9, 64, 7]. We begin by defining convex sets which is the most important object in the convex analysis.

Definition 1. Set $\Omega \subseteq \boldsymbol{R}^d$ is called convex if for any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and $\lambda \in (0, 1)$, we have that $\lambda x + (1 - \lambda)y \in \Omega$.

Convex functions are basically those whose epigraphs (the set of points on or above the graph of the function) are convex sets.

Definition 2. Let $\Omega \subseteq \boldsymbol{R}^d$ be a convex set. We say $f : \Omega \to \boldsymbol{R}^d$ is a convex function if for every $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and any $\lambda \in (0, 1)$, the following is true

$$f\left(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\right) \leq \lambda f(\boldsymbol{x}) + (1-\lambda)\, f(\boldsymbol{y}).$$

Example 3. (Logistic function) The following function defined on the real line is convex. This function is smooth, *i.e.* differentiable of any order.

$$f(t) := \log\left(1 + \exp(-t)\right)$$

Also, we have that $\lim_{t\to+\infty} f(t) = 0$ and $\lim_{t\to-\infty} f(t) = +\infty$.

Example 4. (Hinge function) The following function defined on the real line is convex. It is worth noting that this function is only continuous and not differentiable.

$$h(t) := \max(0, 1-t).$$

Again we have that $\lim_{t\to+\infty} h(t) = 0$ and $\lim_{t\to-\infty} h(t) = +\infty$.

We next define global or local minimizer for a given function f.

Definition 3. For a function $f : \Omega \to \boldsymbol{R}^d$, $\boldsymbol{x}^* \in \Omega$ is called a local minimum if there exists $r > 0$ such that for every $\boldsymbol{y} \in \Omega \cap B_r(\boldsymbol{x}^*)$, it holds that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{y})$. Moreover, $\boldsymbol{x}^* \in \Omega$ is called a global minimum if for every $\boldsymbol{y} \in \Omega$, it holds that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{y})$.

In this thesis, we only deal with unconstrained optimization problems. Because of this, in the rest of this section, we let $\Omega = \boldsymbol{R}^d$.

The main goal of optimization is to analyze the set of minimizers of a given function. Since optimization algorithms only perform iterates based on local information, naturally, we do not expect them to converge to the global optimum. However, for the class of convex functions every local minimum is also global and that is why convex functions are very important. The following theorem lies at the heart of mathematical optimization.

Theorem 1. *For any convex function $f : \boldsymbol{R}^d \to \boldsymbol{R}$, every local minimum is also global.*

Convex functions are not necessary differentiable as in Example 4 above. In such cases, the sub-gradients play the same role as the gradients:

Definition 4. (Subdifferential) Every vector $\boldsymbol{v}$ satisfying the following inequality is called a sub-gradient for the function $f : \boldsymbol{R}^d \to \boldsymbol{R}$ at $\boldsymbol{x}$.

$$f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle \leq f(\boldsymbol{y}), \quad \forall \boldsymbol{y} \in \boldsymbol{R}^d.$$

For all points inside the interior of domain of f, the subdifferential (set of subgradients at a fixed point) denoted by $\partial f(\boldsymbol{x})$ is nonempty, *e.g.* Theorem 3.1.8 [9]. It is worth noting that when f is differentiable at $\boldsymbol{x}$ then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$.

Example 5. Let h be the hinge function from Example 4. Then h is differentiable at any given point except at 0. Sub-differential at 0 are computed as follows.

$$\partial h(0) = [-1, 0].$$

Theorem 2. *(First-Order Necessary Condition) Let $f : \boldsymbol{R}^d \to \boldsymbol{R}$ be a continuously differentiable function. If $\boldsymbol{x}^*$ is a local minimizer of f, it then holds that $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. Conversely, if in addition f is convex, then $\boldsymbol{x}^*$ is a local minimizer of f if and only if $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$.*

Proof. See *e.g.* Theorem 2.2 [66]. □

Theorem 2 leads to the following definition.

Definition 5. Every point $\boldsymbol{x}$ satisfying $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ is called a critical point for f.

Smoothness and strong convexity are defined next.

Definition 6. We say a convex differentiable function f is L-smooth if the following inequality holds for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}^d$.

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{2.1}$$

The bound in (2.1) provides a global upper estimate for f at a given point $\boldsymbol{y}$. The analogous lower bound is called *strong convexity*.

Definition 7. We say a differentiable convex function f is ℓ-strongly convex if the following inequality holds for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}^d$.

$$f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\ell}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq f(\boldsymbol{y}).$$

Algorithm 6: Gradient Descent Algorithm

initialize: $\boldsymbol{x}_0 \in \boldsymbol{R}^d$, $\alpha > 0$
for $k = 0, 1, \cdots$
    Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)$
    $k \leftarrow k + 1$
end

It can be readily verified that every differentiable convex function is 0-strongly convex, this is known as sub-gradient inequality. However, when we discuss strongly convex functions, we only refer to the case where $\ell > 0$. We next state the most basic result regarding the convergence of the GD algorithm.

Theorem 3 (Theorem 3.3, [14]). *Let f be a convex L-smooth function and set $\alpha = \frac{1}{L}$. Assume that the sequence $\{\boldsymbol{x}_k\}_{k=0}^{+\infty}$ is generated by Algorithm 6. The following bound then holds.*

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{k}. \tag{2.2}$$

We conclude this section by an illustrative example of the gradient descent algorithm applied to the logistic function.

Example 6. Consider the following optimization problem:

$$\min_{x \in \boldsymbol{R}} f(x) := \log\left(1 + \exp(-x)\right).$$

We initialize GD at $x_0 = 0$. Then the update formula with $\alpha = 1$ is as follows.

$$x_{k+1} = x_k + \frac{1}{1 + \exp(x_k)}.$$

It can be readily verified that $x_k \approx log(k)$ and $f(x_k) \approx \frac{1}{k}$. In particular, $x_k \to +\infty$ and $f(x_k) \to \inf_{x \in \boldsymbol{R}} f(x)$.

2.1.2 Convergence of SGD

In Chapter 1, we wrote about stochastic optimization in the context of supervised learning. Nonetheless, as uncertainty appears almost in any type of application, stochastic

optimization methods are the sensible choice across a larger variety of problems. Stochastic optimization problems are generally formulated as follows.

$$f(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{P}}\left[F(\boldsymbol{x},\boldsymbol{\zeta})\right]. \tag{2.3}$$

Here $\boldsymbol{x} \in \boldsymbol{R}^d$ is called the decision variable and $\boldsymbol{\zeta}$ is a random variable which follows some distribution $\mathcal{P}$. It is important to underline that the analysis of the SGD algorithm and also its variants are mostly done using this general formulation. Algorithm 9 describes SGD algorithm applied to (2.3).

Algorithm 7: Stochastic gradient descent algorithm

initialize: $\boldsymbol{x}_0 \in \boldsymbol{R}^d, \alpha > 0$
for $k = 0, 1, \cdots$
Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla_{\boldsymbol{x}_k} F(\boldsymbol{x}_k, \boldsymbol{\zeta}_k)$ where $\boldsymbol{\zeta}_k \sim \mathcal{P}$
$k \leftarrow k + 1$
end

Under reasonable assumptions, we can always assume that

$$\mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{P}}\left[\nabla_{\boldsymbol{x}} F(\boldsymbol{x},\boldsymbol{\zeta})\right] = \nabla f(\boldsymbol{x}), \tag{2.4}$$

where the expected value and the derivative are interchanged. In view of (2.4), $\nabla_{\boldsymbol{x}} F(\boldsymbol{x},\boldsymbol{\zeta})$ are called *unbiased stochastic gradients*. For the convergence analysis of Algorithm 7, we need to assume that the stochastic gradients $\nabla_{\boldsymbol{x}} F(\boldsymbol{x},\boldsymbol{\zeta})$ satisfy some uniformity when $\boldsymbol{\zeta} \sim \mathcal{P}$. The following condition is commonly assumed for the convergence analysis of the Algorithm 7.

$$\mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{P}}\left[\|\nabla f(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} F(\boldsymbol{x},\boldsymbol{\zeta})\|^2\right] \le \tau^2 \tag{2.5}$$

In words, stochastic gradients $\nabla_{\boldsymbol{x}} F(\boldsymbol{x},\boldsymbol{\zeta})$ are expected to lie close to the main gradient $\nabla f(\boldsymbol{x})$. We have the following result.

Theorem 4 (Theorem 4.1, [50]). *Suppose that the regularity conditions* (2.4) *and* (2.5) *hold. Then with the step-size* $\alpha = \frac{1}{L+\frac{1}{\beta}}$ *where* $\beta = \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|}{\tau} \cdot \sqrt{\frac{2}{k}}$ *the following holds.*

$$\mathbb{E}\left[f\left(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{x}_i\right)\right] - f(\boldsymbol{x}^*) \le \|\boldsymbol{x}_1 - \boldsymbol{x}^*\| \cdot \tau \cdot \sqrt{\frac{2}{k}} + \frac{L\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{k}. \tag{2.6}$$

Notice the difference between the bounds in (2.2) and (2.6). Indeed, all the stochastic algorithms only optimize the expected performance of the model rather than providing

deterministic guarantees. It is known that the bound in Theorem 4 is optimal. In fact, from the classical theory of convex programming [63, 64], we know that for a stochastic first order method where at each iteration an unbiased stochastic gradient is evaluated, finding an ϵ-solution *i.e.* a point $\boldsymbol{x}_\epsilon$ such that

$$\mathbb{E}\left[f(\boldsymbol{x}_\epsilon)\right] - \min f \leq \epsilon,$$

requires $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ or, $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ if f is strongly convex, stochastic gradient evaluations.

It is worth mentiong that SGD variants have been the default algorithms of choice across many machine learning applications. In particular, adaptive gradient methods have gained paramount popularity in training deep neural networks where their key feature is that they apply a preconditioning matrix for the gradient updates at each iteration. In other words, different learning rates are considered for different coordinates in the gradient update. In practice diagonal preconditioning are used whereas theoretical results are based on full-matrix preconditioning. Examples of adaptive gradient methods include [17, 82, 88, 48, 74].

2.2 Probability theory

In this section, we establish the basic notation and record some preliminary results from probability theory that we will use throughout the thesis.

2.2.1 Probability distributions

A random variables is a variable that can take on different values randomly.

Example 7. Let Ω be a set and $A \subseteq \Omega$ be a random subset of Ω. The indicator function of the subset A is defined as follows.

$$1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A. \end{cases}$$

A description of how likely a random variable is to take on each of its possible states is called a probability distribution. For example, for a given real-valued random variable X, we say X has density function p whenever it holds that

$$\mathbb{P}\left(a \leq X \leq b\right) = \int_a^b p(x)dx. \tag{2.7}$$

To be a probability distribution, a function p needs to satisfy the following properties:

1. $p(x) \geq 0$ for all $x \in \boldsymbol{R}$.
2. $\int_{-\infty}^{+\infty} p(x)dx = 1$.

When (2.7) holds, we write

$$X \sim \mathbb{P}. \tag{2.8}$$

The probability density functions for multi-variate random variables are defined similarly as in (2.7). In other words, for a random variable X in $\boldsymbol{R}^d$, we say that X follows the probability distribution $\mathbb{P}$ if

$$\mathbb{P}\left(X \in [a_1, b_1] \times \cdots [a_d, b_d]\right) = \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} p(x_1, \cdots, x_d) dx_1 \cdots dx_d.$$

Assuming that p satisfies (2.7), the cumulative distribution function of the random variable X is defined as follows.

$$F(b) := \mathbb{P}\left(X \leq b\right) = \int_{-\infty}^{b} p(t)dt.$$

It can be easily verified that

$$p(t) = \frac{d}{dt} F(t).$$

The expected value of some function $f(X)$ with respect to a probability distribution $p(x)$ is the average value that f takes on when $X \sim \mathbb{P}$ as in (2.8). In other words, we define

$$\mathbb{E}_{X \sim \mathbb{P}}\left[f(X)\right] := \int p(x) f(x) dx.$$

Example 8. The probability density function of a random variable X which is uniformly distributed on the interval $[a, b]$ is given by

$$p(x) := \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b]. \end{cases} \tag{2.9}$$

Therefore, it holds that

$$\mathbb{E}\left[X\right] = \int_a^b t dt = \frac{b-a}{2}.$$

When a random variable X follows the uniform distribution on $[a, b]$, it is written $X \sim U(a, b)$.

2.2.2 Normal distributions

Gaussian distributions, also known as normal distributions, are the most commonly used distribution over real numbers. The probability density function of a univariate Gaussian with mean μ and variance σ^2 is described by:

$$\varphi_{\mu,\sigma^2}(t) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right).$$

In particular, we denote a random variable ξ distributed as a Gaussian with mean μ and variance σ^2 by $\xi \sim N(\mu, \sigma^2)$ to mean $\mathbb{P}(\xi \leq t) = \int_{-\infty}^{t} \varphi_{\mu,\sigma^2}(t)\, dt$. When the random variable $\xi \sim N(0,1)$, we denote its cumulative density function as

$$\Phi(t) := \mathbb{P}(\xi \leq t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{s^2}{2}\right) ds,$$

and its complement by $\Phi^c(t) = 1 - \Phi(t)$. The symmetry of a normal around its mean yields the identity, $\Phi(t) = \Phi^c(-t)$.

One can, analogously, formulate a higher dimensional version of the univariate normal distribution called a *multivariate normal distribution*. A random vector is a multivariate normal distribution if every linear combination of its component is a univariate normal distribution. We denote such multivariate normals by $\boldsymbol{\xi} \sim N(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} \in \boldsymbol{R}^d$ and Σ is a symmetric positive semidefinite $d \times d$ matrix.

Normal distributions have interesting properties which simplify our computations throughout Chapter 3. We list those which we specifically rely on. See [21] for proofs. Below, $\boldsymbol{v}, \boldsymbol{v}' \in \boldsymbol{R}^d$, $r \in \boldsymbol{R}$, $\boldsymbol{\xi} \sim N(\boldsymbol{\mu}, \sigma^2 I_d)$, $\xi \sim N(\mu, \sigma^2)$, and $\psi \sim N(0,1)$.

Fact 4. Consider random variables of the form $\boldsymbol{v}^T\boldsymbol{\xi} + r$, i.e. affine functions of a given normal distribution. A fundamental property of normal distributions is that they stay in the same class of distributions after any such transformation. In particular, it holds that

$$\boldsymbol{v}^T\boldsymbol{\xi} + r \sim N(\boldsymbol{v}^T\boldsymbol{\mu} + r, \sigma^2\|\boldsymbol{v}\|^2). \tag{2.10}$$

Fact 5. Working with independent random variables makes the analysis significantly easier. In particular, it is essential for us to know when the two random variables $\boldsymbol{v}^T\boldsymbol{\xi}$ and $\boldsymbol{v}'^T\boldsymbol{\xi}$ are independent. We will use the following simple fact multiple times in Chapter 3:

$$\boldsymbol{v}^T\boldsymbol{\xi} \text{ and } \boldsymbol{v}'^T\boldsymbol{\xi} \text{ are independent} \quad \text{if and only if} \quad \boldsymbol{v}^T\boldsymbol{v}' = 0. \tag{2.11}$$

Fact 6. Truncated normal distribution appear in our analysis. We will need the following simple fact:

$$\mathbb{E}_\xi[\xi 1_{\{\xi \leq b\}}] = 0 \Longrightarrow \Phi\left(\frac{b-\mu}{\sigma}\right) \cdot \exp\left(\frac{1}{2} \cdot \left(\frac{b-\mu}{\sigma}\right)^2\right) - \frac{\sigma}{\mu} = 0. \tag{2.12}$$

Fact 7. We conclude our remarks on normal distributions with the statement of two facts about the expected value of their norm. The following hold:

$$\mathbb{E}\left[\|\boldsymbol{\xi}\|^2\right] = \|\boldsymbol{\mu}\|^2 + d\sigma^2, \quad \mathbb{E}_\xi[|\xi|] \leq \sqrt{\frac{2}{\pi}} \cdot \sigma + |\mu| \quad \text{and} \quad \mathbb{E}\left[|\psi|\right] = \sqrt{\frac{2}{\pi}}. \tag{2.13}$$

2.2.3 Martingales and stopping times

Here we state some relevant definitions and theorems regarding Martingales and stopping times used in Chapters 3 and 4. We refer the reader to [18] for further details.

2.2.4 Martingales

For any probability space, $(\mathbb{P}, \Omega, \mathcal{F})$, we call a sequence of σ-algebras, $\{\mathcal{F}_k\}_{k=0}^\infty$, a *filtration* provided that $\mathcal{F}_i \subset \mathcal{F}$ and $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq 2^\Omega$ holds. Given a filtration, it is natural to define a sequence of random variables $\{X_k\}_{k=0}^\infty$ with respect to the filtration, namely X_k is a $\mathcal{F}_k$-measurable function. If, in addition, the sequence satisfies

$$\mathbb{E}[|X_k|] < \infty \quad \text{and} \quad \mathbb{E}[X_{k+1}|\mathcal{F}_k] \leq X_k \quad \text{for all } k, \tag{2.14}$$

we say $\{X_k\}_{k=0}^\infty$ is a *supermartingale*. Similarly, if the right hand side inequality in (2.14) holds with an equality, then we refer to $\{X_k\}_{k=0}^\infty$ as a *martingale*. *Submartingales* are defined in a similar fashion. Note that martingales, supermartingales and submartingales are the stochastic analogous of monotonic sequences. In other words, a sequence of random variables is called a supermartingale, submartingale and martingale if *in expectation* it is a non-increasing, non-decreasing and constant real-valued sequence respectively. As for bounded monotonic sequences, bounded supermartingales, submartingales or martingales converge almost surely; see *e.g.* Theorem 27.1 [45].

2.2.5 Stopping times

In probability theory, we are often interested in the (random) time at which a given stochastic sequence exhibits a particular behavior. Such random variables are known as *stopping times*. Precisely, a stopping time is a random variable $T : \Omega \to \mathbb{N} \cup \{0, \infty\}$ where the event $\{T = k\} \in \mathcal{F}_k$ for each k, i.e., the decision to stop at time k must be measurable with respect to the information known at that time. As we illustrate in Chapter 3, a connection between stopping criteria (i.e. the decision to stop an algorithm) and stopping times naturally exists.

Example 9. (Random walk on $\mathbb{Z}$) A random walk is an stochastic process formed by iteratively summing independent, identically distributed random variables. Random walks on lattices for instance has been extensively studied [80] and its simplest case is constructed as the following stochastic process: Suppose that $\{X_k\}_{k=0}^{+\infty}$ is an iid sequence of Bernoulli random variables with $\mathbb{P}(X_k = 1) = \frac{1}{2}$. Denote by $S_n = X_1 + \cdots + X_n$ for every positive integer n. For every positive integer K, define the random variable T_K as follows.

$$T_K = \inf\{n : S_n \geq K\}.$$

Clearly, T_K is a stopping time with respect to the filtration $\mathcal{F}_k := \sigma(X_1, \cdots, X_k)$.

Supermartingales and stopping times are closely tied together, as seen in the theorem below, which gives a bound on the expectation of a stopped supermartingale.

Theorem 5 (See [18] Theorem 4.8.5)**.** *Suppose that $\{X_k\}_{k=0}^{\infty}$ is a supermartingale w.r.t to the filtration $\{\mathcal{F}_k\}_{k=0}^{\infty}$ and let T be any stopping time satisfying $\mathbb{E}[T] < \infty$. Moreover if $\mathbb{E}[|X_{k+1} - X_k||\mathcal{F}_k] \leq B$ a.s. for some constant $B > 0$, then it holds that $\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$.*

Tower Rule Tower Rule is a simple lemma that we will use in our analysis multiple times.

Lemma 1 (Tower Rule)**.** *Let $(\mathbb{P}, \Omega, \mathcal{F})$ be a probability space with two sub σ-algebras $\mathcal{G}_1 \subseteq \mathcal{G}_2$. Given a random variable X on this space, if $\mathbb{E}[|X|] < +\infty$, then the following holds.*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] = \mathbb{E}[X|\mathcal{G}_1]. \tag{2.15}$$

2.2.6 Concentration inequality

Non-asymptotic concentration bounds have been the subject of intensive study in the data science literature [86]. This is mostly due to the fact that stochastic algorithms in

optimization have risen to an unprecedented popularity. A non-asymptotic concentration inequality asserts that a random variable X concentrates around its mean *i.e.* $\mathbb{E}[X]$, with high probability. In contrast with other types of statistical analysis, the non-asymptotic point of view does not require the dimension to go off to infinity. A classic example of asymptotic results is the law of large numbers. We state and prove Hoeffding's and Azuma's concentration inequality. We need Azuma's inequality in Chapter 4.

2.2.7 Hoeffding's inequality

Hoeffding's inequality illustrates a concentration bound for the sum of uniformly bounded random variables. We state and prove it below.

Lemma 2 (Hoeffding's inequality)**.** *Let $X_1, \cdots, X_N$ be random variables such that $X_k \in [a_k, b_k]$ almost surely for all $k \in [N]$. Denote by $S_N := X_1 + \cdots + X_N$. The following bound then holds.*

$$\mathbb{P}\left(S_N - \mathbb{E}\left[S_N\right] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{k=1}^{N}\left(b_k - a_k\right)^2}\right). \tag{2.16}$$

Proof of Lemma 2 follows from Hoeffding's lemma which provides an upper bound for the moment generating function of a bounded random variable. We state and prove Hoeffding's lemma below [1].

Lemma 3. *Suppose that X is a random variable such that $X \in [a, b]$ a.s. The following bound then holds for all $s \in \boldsymbol{R}$.*

$$\mathbb{E}\left[\exp\left(s\left(X - \mathbb{E}\left[X\right]\right)\right)\right] \leq \exp\left(\frac{1}{8}s^2\left(b-a\right)^2\right). \tag{2.17}$$

Proof. First note that without loss of generality we can assume that $\mathbb{E}\left[X\right] = 0$. Denote $p = \frac{-a}{b-a}$ and define $L(x) := -px + \ln\left(1 - p + pe^x\right)$. We have that

$$e^{L(x)} = \left(1 - p + pe^x\right)e^{-px} = (1-p)e^{-px} + pe^{(1-p)x}. \tag{2.18}$$

Plugging in $x^* = s(b-a)$ into (2.18), we will obtain that

$$e^{L(x^*)} = (1-p)e^{sa} + pe^{sb} = \frac{b - \mathbb{E}[X]}{b-a}e^{sa} + \frac{\mathbb{E}[X] - a}{b-a}e^{sb} \geq \mathbb{E}\left[e^{sX}\right]. \tag{2.19}$$

Here the last inequality follows from convexity of the exponential function. Now note that $L(0) = L'(0) = 0$ and $L''(x) \leq \frac{1}{4}$ for all x. Therefore, it holds that $L(x) \leq \frac{x^2}{8}$ and hence using (2.19), we will conclude the proof. □

We are now ready to prove Hoeffding's inequality Lemma 2.

Proof of Lemma 2. By Markov inequality, for any $s > 0$, we will have that

$$\begin{aligned}
\mathbb{P}\left(S_N - \mathbb{E}\left[S_N\right] \geq \epsilon\right) &= \mathbb{P}\left(\exp\left(s\left(S_N - \mathbb{E}\left[S_N\right]\right)\right) \geq e^{s\epsilon}\right) \\
&\leq e^{-s\epsilon}\mathbb{E}\left[\exp\left(s\left(S_N - \mathbb{E}\left[S_N\right]\right)\right)\right] \\
&= e^{-s\epsilon}\prod_{k=1}^{N}\mathbb{E}\left[\exp\left(s\left(X_k - \mathbb{E}\left[X_k\right]\right)\right)\right] \\
&\leq e^{-s\epsilon}\prod_{k=1}^{N}\exp\left(\frac{s^2}{8}\left(b_k - a_k\right)^2\right) \\
&= e^{-s\epsilon}\exp\left(\frac{s^2}{8}\sum_{k=1}^{N}\left(b_k - a_k\right)^2\right).
\end{aligned}$$

Letting $s = \frac{4\epsilon}{\sum_{k=1}^{N}(b_k - a_k)^2}$ and using the above chain of bounds, we will conclude the lemma. □

2.2.8 Azuma's inequality

We now state and prove Azuma's inequality. This concentration inequality provides high probability concentration bounds for Martingales with bounded differences.

Lemma 4. *(Azuma's Inequality) Suppose that $\{X_k\}_{k=0}^{+\infty}$ is a martingale that satisfies $|X_k - X_{k-1}| \leq c_k$ almost surely for all $k \geq 1$. Here $\{c_k\}_{k=0}^{+\infty}$ is a sequence of positive real numbers. Then for all positive integer N and any $\epsilon > 0$, the following bound is true.*

$$\mathbb{P}\left(|X_N - X_0| \geq \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^{N} c_k^2}\right).$$

Proof. We first show that the following holds.

$$\mathbb{P}\left(X_N - X_0 \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^{N} c_k^2}\right). \tag{2.20}$$

Note that for $s > 0$, we have that

$$\begin{aligned}
\mathbb{P}\left(X_N - X_0 \geq \epsilon\right) &= \mathbb{P}\left(\exp\left(s(X_N - X_0)\right) \geq \exp(s\epsilon)\right) \\
&\leq \exp(-s\epsilon)\mathbb{E}\left[\exp\left(s(X_N - X_0)\right)\right] \\
&= \exp(-s\epsilon)\mathbb{E}\left[\mathbb{E}\left[\exp\left(s(X_N - X_0)\right)|X_1, \cdots, X_{N-1}\right]\right] \\
&= \exp(-s\epsilon)\mathbb{E}\left[\exp\left(s(X_N - X_{N-1})\right)\mathbb{E}\left[\exp\left(s\sum_{k=1}^{n-1}(X_k - X_{k-1})\right)|X_1, \cdots, X_{N-1}\right]\right] \\
&\leq \exp(-s\epsilon)\mathbb{E}\left[\exp\left(s\sum_{k=1}^{n-1}(X_k - X_{k-1})\right)\mathbb{E}\left[\exp\left(s(X_N - X_{N-1})\right)|X_1, \cdots, X_{N-1}\right]\right] \\
&\leq \exp\left(-s\epsilon + \frac{s^2 c_N^2}{8}\right)\mathbb{E}\left[\left[\exp\left(s\sum_{k=1}^{n-1}(X_k - X_{k-1})\right)\right]\right].
\end{aligned}$$

Here the second equality follows from tower rule Lemma 1 and the last inequality follows from Hoeffding's inequality. Iterating, we therefore obtain that

$$\mathbb{P}\left(X_N - X_0 \geq \epsilon\right) \leq \exp\left(-s\epsilon + \frac{s^2\sum_{k=1}^N c_k^2}{8}\right) \tag{2.21}$$

Plugging in $s = \frac{4\epsilon}{\sum_{k=1}^N c_k^2}$ in (2.21), we will obtain (2.20). Similarly, we will obtain

$$\mathbb{P}\left(X_N - X_0 \leq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^N c_k^2}\right). \tag{2.22}$$

Combining (2.20) and (2.22), we will conclude the lemma. □

2.2.9 Concentration for norm

Our last concentration bound is about the norm of Gaussian variables essentially asserts that the norm of Gaussian random variables concentrate around its mean with overwhelming probability.

Lemma 5. *Let $\boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2 I_d)$. The following bound always hold.*

$$\mathbb{P}\left(\|\boldsymbol{\xi}\| \geq \epsilon\sigma + \mathbb{E}\left[\|\boldsymbol{\xi}\|\right]\right) \leq \exp\left(-\frac{\epsilon^2}{2}\right) \quad \textit{for all } \epsilon \geq 0. \tag{2.23}$$

In addition, the following is true

$$\mathbb{E}\left[\|\boldsymbol{\xi}\|\right] \leq \sigma\sqrt{\frac{d+1}{4}}. \tag{2.24}$$

Proof. See *e.g.* [86], p.40, Theorem 2.26. It is also well-known that $\mathbb{E}\left[\|\boldsymbol{\xi}\|\right] = \frac{\sigma\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}$. Now by Gautschi's inequality, see [23], it follows that for all $d > 1$,

$$\mathbb{E}\left[\|\boldsymbol{\xi}\|\right] \leq \sigma\sqrt{\frac{d+1}{4}}. \tag{2.25}$$

The proof is complete. □

2.2.10 Markov Chain Theory

A Markov chain is a stochastic system such that the future state does not depend on how the system has arrived at the current state. For example, stochastic iterative algorithms which we discussed in Chapter 1 can be cast as Markov chains. An important concept in the Markov chain theory is the idea of drift analysis. For a comprehensive treatment of Markov chain see *e.g.* [56].

2.2.11 Drift criterion

Consider a Markov chain $\{\boldsymbol{\theta}_k\}_{k=0}^{+\infty}$ where for some set C and some non-negative function V, the expected value of $V(\boldsymbol{\theta}_k)$ decreasing by a fixed constant whenever $\boldsymbol{\theta}_{k-1} \notin C$, *e.g.*

$$\mathbb{E}\left[V(\boldsymbol{\theta}_k)|\boldsymbol{\theta}_{k-1}\right] \leq V(\boldsymbol{\theta}_{k-1}) - 1 \quad \text{whenever } \boldsymbol{\theta}_{k-1} \notin C. \tag{2.26}$$

Intuitively, the chain drifts towards set C, meaning that whenever the current iterate lies outside of C, it tends to move back towards C. Having this drift criterion at hand, we can establish a bound for the expected value of the first time that the chain $\{\boldsymbol{\theta}_k\}_{k=0}^{+\infty}$ lies inside C. This idea which lies at the heart of our analysis in Chapter 3 is formulated below.

Proposition 1 ([56], Theorem 11.3.4)**.** *Given a Markov chain $\{\boldsymbol{\theta}_k\}_{k=0}^{+\infty}$ with $\boldsymbol{\theta}_k \in \boldsymbol{R}^d$, assume that there exist a non-negative function V and a subset $C \subseteq \boldsymbol{R}^d$ such that the drift criterion (2.26) holds. Denote by τ_1 the smallest positive k that $\boldsymbol{\theta}_k \in C$. The following bound is then true.*

$$\mathbb{E}\left[\tau_1|\boldsymbol{\theta}_0\right] \leq V(\boldsymbol{\theta}_0).$$

Proof. Fix a positive integer n. The following identity holds

$$V(\boldsymbol{\theta}_{\tau_1\wedge n}) = V(\boldsymbol{\theta}_0) + \sum_{k=1}^{n} \left[V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k-1})\right] 1_{\{\tau_1\wedge n\geq k\}}.$$

To ease the notation, we denote $\mathcal{F}_{-1} = \sigma\left(\{\boldsymbol{\theta}_0 = \boldsymbol{\theta}\}\right)$. We continue

$$\begin{aligned}\mathbb{E}\left[V(\boldsymbol{\theta}_{\tau_1\wedge n})|\mathcal{F}_{-1}\right] &= V(\boldsymbol{\theta}) + \sum_{k=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\left(V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k-1})\right) 1_{\{\tau_1\wedge n\geq k\}}|\mathcal{F}_{k-1}\right]|\mathcal{F}_{-1}\right] \\ &= V(\boldsymbol{\theta}) + \mathbb{E}\left[\left(\sum_{k=1}^{n} \mathbb{E}\left[V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k-1})|\mathcal{F}_{k-1}\right] 1_{\{\tau_1\wedge n\geq k\}}\right)|\mathcal{F}_{-1}\right] \quad (2.27)\end{aligned}$$

We now upper estimate the quantity inside the bracket in the above equation. Using (2.26), since $\boldsymbol{\theta}_{k-1} \notin C$ for all $1 \leq k \leq \tau_1$, we have

$$\mathbb{E}\left[V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k-1})|\mathcal{F}_{k-1}\right] 1_{\{\tau_1\wedge n\geq k\}} \leq -1_{\{\tau_1\wedge n\geq k\}}. \quad (2.28)$$

Plugging in the estimate (2.28) into (2.27), we obtain

$$\sum_{k=1}^{n} \mathbb{E}\left[\left(V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k-1})\right) 1_{\{\tau_1\wedge n\geq k\}}|\mathcal{F}_{k-1}\right] \leq -\mathbb{E}\left[\tau_1 \wedge n\right].$$

Therefore we have

$$\mathbb{E}\left[V(\boldsymbol{\theta}_{\tau_1\wedge n})|\mathcal{F}_{-1}\right] \leq V(\boldsymbol{\theta}) - \mathbb{E}\left[\tau_1 \wedge n|\mathcal{F}_{-1}\right].$$

Since $0 \leq \mathbb{E}\left[V(\boldsymbol{\theta}_{\tau_1\wedge n})|\mathcal{F}_{-1}\right]$, this gives $\mathbb{E}\left[\tau_1 \wedge n|\mathcal{F}_{-1}\right] \leq V(\boldsymbol{\theta})$. By monotone convergence theorem the claim follows. □

Chapter 3

A Termination Criterion for SGD for Binary Classification

In this chapter, the binary linear classification problem is considered where a linear classifier is sought by using the SGD algorithm to minimize the logistic and hinge expected loss functions. In our setting, we observe a sequence of data points $\{(\boldsymbol{\zeta}_k, y_k)\}_{k=0}^{+\infty}$ where $y_k \in \{0, 1\}$. Letting $\{\boldsymbol{\theta}_k\}_{k=0}^{+\infty}$ be a sequence generated by the SGD algorithm, we

$$\text{Terminate when } (2y_{k+1} - 1)\boldsymbol{\zeta}_{k+1}^T \boldsymbol{\theta}_k \geq 1, \tag{3.1}$$

Notice that the termination criterion (3.1) implies that the iterate $\boldsymbol{\theta}_k$ is making a large enough margin with the data point $(\boldsymbol{\zeta}_{k+1}, y_{k+1})$. The main results of this chapter [3] are as follows:

- We will analyze the termination criterion (3.1) by assuming that the data is distributed according to a Gaussian mixture model. With this assumption, we show that $\boldsymbol{\theta}_k$ is converging to an optimal classifier as $k \to +\infty$ (Lemma 6).
- Denoting by T the iterate where (3.1) occurs, we will first show that T is finite almost surely (Theorems 6 and 7). Second, provided that the variance σ^2 within the vectors $\boldsymbol{\zeta}_k$ is not too large, we will prove that the expected value of T decays exponentially with respect to σ^2 (Theorem 6).
- We prove that the accuracy of the classifier at termination nearly matches the accuracy of an optimal classifier (Theorem 8). Accuracy is the fraction of predictions

that a classification model got right while an optimal classifier minimizes the probability of misclassification when the sample is drawn from the same distribution as the training data.

- In Section 3.4, we empirically evaluate the performance of our stopping criterion versus a baseline competitor. We compare performances on both synthetic (Gaussian and heavy-tailed t-distribution) as well as real data sets (MNIST [51] and CIFAR-10 [49]). In our experiments, we observe that our test yields relatively accurate classifiers with small variation across multiple runs.

The outline of this chapter is as follows: First, in Section 3.1, we recall the binary classification problem from Chapter 1. We also discuss the Gaussian mixture models. Next, in Section 3.2, we propose a termination criterion for SGD applied to expected logistic and hinge loss functions (Algorithm 8). Next, in Section 3.3, we provide theoretical evidence for the effectiveness of our stopping criterion by considering Gaussian mixture models. In Section 3.4, we conduct numerical experiments on synthetic and real data sets.

3.1 Binary classification problem

We consider the binary classification problem where the data is generated based on a Gaussian mixture model (GMM) [75, 54, 55]. In a Gaussian mixture model, the data is generated based on two Gaussian distributions $\mathcal{P}_0 \sim N(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{P}_1 \sim N(\boldsymbol{\mu}_2, \Sigma_2)$ at each step according to a Bernoulli distribution. In other words, for some fixed $p \in (0,1)$, a data point is generated from $\mathcal{P}_0$ with probability p and from $\mathcal{P}_1$ with probability $1-p$. Here the samples $(\boldsymbol{\zeta}, y) \in \boldsymbol{R}^d \times \{0,1\}$. We consider linear predictors which means that for a fixed suitable loss function ℓ, we have that $\ell_{\boldsymbol{\theta}}(\boldsymbol{\zeta}, y) = \ell(\boldsymbol{\theta}^T \boldsymbol{\zeta}, y)$.

We consider logistic and hing loss function (Examples 3 and 4). Our main goal here is to compute the exact solution to the expected loss functions when GMM is considered, see Lemma 6. We recall that in logistic regression the loss function is defined as follows

$$\ell(x, y) := -yx + \log\left(1 + \exp(x)\right). \tag{3.2}$$

Also, the hinge loss is defined as the following

$$\ell(x, y) := \begin{cases} \max(1-x, 0) & y = 1, \\ \max(1+x, 0) & y = 0. \end{cases} \tag{3.3}$$

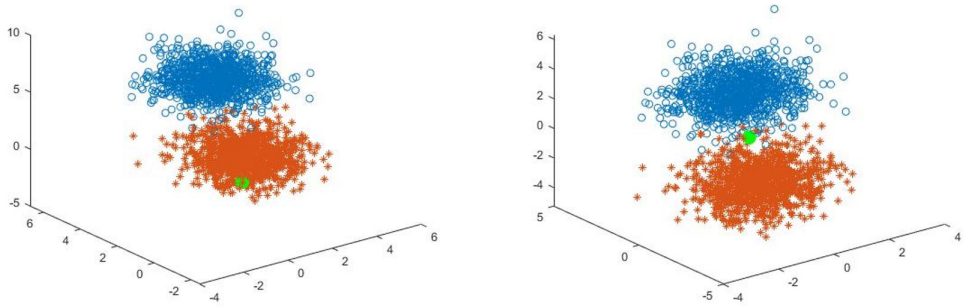

Figure 3.1: Re-centring phase. Here synthetic Gaussian data is generated in $\boldsymbol{R}^3$ and the green dot denotes the origin.

We thus analyze learning by minimizing an expected loss problem of linear predictors (*i.e.*, without bias) of the form

$$\mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}}[\ell(\boldsymbol{\zeta}^T\boldsymbol{\theta}, y)]$$

using logistic and hinge regression. Further, we will simplify our argument via two preliminary steps. First, we re-centre the data points using some preliminary samples, *i.e.* $\boldsymbol{\mu}_0 = -\boldsymbol{\mu}_1$. We enforce this assumption, with minimal loss in accuracy, by recentering the data using a preliminary round of sampling. See Fig 3.1. Next because of the homogeneity assumption (*i.e.* data is origin-centered), we can simplify the notation by redefining our training examples to be $\boldsymbol{\xi}_k := (2y_k - 1)\boldsymbol{\zeta}_k$ and then assuming that for all $k \geq 0$, $y_k = 1$. Then the new samples $\boldsymbol{\xi}$ can be drawn from a *single*, mixed distribution $\mathcal{P}_*$ with mean $\boldsymbol{\mu} := \boldsymbol{\mu}_1$ where sampling $\boldsymbol{\xi} \sim \mathcal{P}_1$ occurs with probability 0.5 and $-\boldsymbol{\xi} \sim \mathcal{P}_0$ occurs with probability 0.5. For simplicity, we assume that $\Sigma_1 = \Sigma_2 = \sigma^2 I_d$. Therefore, we obtain that $\mathcal{P}_* \sim N(\boldsymbol{\mu}, \sigma^2 I_d)$.

We make this simplification and, from this point on, we analyze the following optimization problem:

$$\min_{\boldsymbol{\theta}\in\boldsymbol{R}^d} f(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{P}_*}[\ell(\boldsymbol{\xi}^T\boldsymbol{\theta}, 1)]. \tag{3.4}$$

Let us remark that the right-hand side of (3.4) is differentiable with respect to $\boldsymbol{\theta}$ in either cases of logistic and hinge loss functions. Indeed, in case of hinge loss, note that for any

$\boldsymbol{\theta}_{k-1}$, the function $\boldsymbol{\xi}_k \mapsto \ell(\boldsymbol{\xi}_k^T \boldsymbol{\theta}_{k-1}, 1)$ is almost surely differentiable as $\mathbb{P}_{\boldsymbol{\xi}_k}\left(\boldsymbol{\xi}_k^T \boldsymbol{\theta}_{k-1} = 1\right) = 0$. Hence, we consider the expectation in (3.4) to be over $\mathbb{R}^d \backslash \left\{\boldsymbol{\xi}_k : \boldsymbol{\xi}_k^T \boldsymbol{\theta}_{k-1} = 1\right\}$ on which the argument is differentiable with respect to $\boldsymbol{\theta}_{k-1}$. In the lemma below, we provide closed-from formula for the minimizer of the expected loss where $\mathcal{P}_* \sim N(\boldsymbol{\mu}, \sigma^2 I_d)$. We will use this result in Chapter 3.

Lemma 6 (Minimizer of the logistic and hinge expected loss)**.** *The function f defined in (3.4) with ℓ defined in (3.2) or (3.3) has a unique minimizer at $\boldsymbol{\theta}^* = \rho^* \boldsymbol{\mu}$ for some $\rho^* \in (0, +\infty)$. Moreover, let $r = \rho^* \sigma^2$. Then in the case of logistic regression, it holds that $r = 2$ and in the case of hinge loss, $w = \frac{\sigma}{r\|\boldsymbol{\mu}\|} - \frac{\|\boldsymbol{\mu}\|}{\sigma}$ satisfies*

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{\sigma}{\|\boldsymbol{\mu}\|} = \Phi(w) \cdot \exp(\tfrac{1}{2} w^2). \tag{3.5}$$

Proof. We consider the logistic and hinge loss case separately.

Logistic loss. We have

$$f(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\xi} \sim N(\boldsymbol{\mu}, \sigma^2 I_d)}[-\boldsymbol{\theta}^T \boldsymbol{\xi} + \log(1 + \exp(\boldsymbol{\theta}^T \boldsymbol{\xi}))].$$

Clearly, f is a convex function. We next observe that for any $\boldsymbol{v}, \boldsymbol{\theta} \in \boldsymbol{R}^d$ with $\boldsymbol{v}^T \boldsymbol{\theta} = 0$, it holds that

$$\boldsymbol{v}^T \nabla f\left(\boldsymbol{\theta}\right) = \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\boldsymbol{\xi}^T \boldsymbol{v}}{1 + \exp(\boldsymbol{\xi}^T \boldsymbol{\theta})}\right] = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\xi}^T \boldsymbol{v}] \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{1 + \exp(\boldsymbol{\xi}^T \boldsymbol{\theta})}\right] = \boldsymbol{v}^T \boldsymbol{\mu} \cdot \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{1 + \exp(\boldsymbol{\xi}^T \boldsymbol{\theta})}\right]. \tag{3.6}$$

Here we used that $\boldsymbol{\xi}^T \boldsymbol{v}$ and $\boldsymbol{\xi}^T \boldsymbol{\theta}$ are independent random variables and the expectation of the product of two uncorrelated random variables is the product of the expectations. Now note that for any $\boldsymbol{\theta}$, the quantity $\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{1+\exp(\boldsymbol{\xi}^T \boldsymbol{\theta})}\right]$ is strictly positive. Therefore, if $\boldsymbol{v}^T \boldsymbol{\theta} = 0$ and $\nabla f(\boldsymbol{\theta}) = \mathbf{0}$ then, using (3.6), we obtain that $\boldsymbol{v}^T \boldsymbol{\mu} = 0$. Hence, we established that $\nabla f(\boldsymbol{\theta}) = \mathbf{0}$ implies $\boldsymbol{\theta} = \rho \boldsymbol{\mu}$ for some $\rho \in \boldsymbol{R}$. On the other hand, using (3.6) again, we have that $\nabla f(\rho \boldsymbol{\mu}) = 0$ if and only if $\boldsymbol{\mu}^T \nabla f(\rho \boldsymbol{\mu}) = 0$. To see the only if direction, suppose $\boldsymbol{\mu}^T \nabla f(\rho \boldsymbol{\mu}) = 0$ and $\nabla f(\rho \boldsymbol{\mu}) \neq 0$. Then we have $\nabla f(\rho \boldsymbol{\mu}) = \boldsymbol{v}$ where the vector $\boldsymbol{v}$ is nonzero such that $\boldsymbol{v}^T \boldsymbol{\mu} = 0$. By (3.6), we deduce $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T \nabla f(\rho \boldsymbol{\mu}) = 0$ yielding a contradiction.

Next, we consider the function,

$$g(\rho) := -\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\boldsymbol{\mu}^T \boldsymbol{\xi}}{1 + \exp(\rho \boldsymbol{\mu}^T \boldsymbol{\xi})}\right].$$

Observe that $g(\rho) = \boldsymbol{\mu}^T \nabla f(\rho\boldsymbol{\mu})$. Therefore, if we can show $g(\rho)$ has a unique zero at $\rho = \frac{2}{\sigma^2} =: \rho^*$, we can conclude that $\boldsymbol{\mu}^T \nabla f(\rho^*\boldsymbol{\mu}) = 0$ which, in turn, gives us that $\rho^*\boldsymbol{\mu}$ is the unique solution to $\nabla f(\rho^*\boldsymbol{\mu}) = 0$. It remains to show that ρ^* is the unique zero of g. By (2.10), $z := \boldsymbol{\mu}^T\boldsymbol{\xi} \sim N(\|\boldsymbol{\mu}\|^2, \sigma^2\|\boldsymbol{\mu}\|^2)$. Therefore, this yields

$$g(\rho) = \frac{1}{\sigma\|\boldsymbol{\mu}\|\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{z}{1+\exp(\rho z)} \exp\left(-\frac{(z-\|\boldsymbol{\mu}\|^2)^2}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right) dz.$$

Expanding out the term inside the integral, we conclude

$$\begin{aligned}\frac{z}{1+\exp(\rho z)} \exp\left(-\frac{(z-\|\boldsymbol{\mu}\|^2)^2}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right) &= \frac{z}{2\cosh\left(\frac{\rho z}{2}\right)} \exp\left(-\frac{\rho z}{2} - \frac{(z-\|\boldsymbol{\mu}\|^2)^2}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right) \\ &= \frac{z}{2\cosh\left(\frac{\rho z}{2}\right)} \exp\left(-\frac{z^2 + (\rho\sigma^2\|\boldsymbol{\mu}\|^2 - 2\|\boldsymbol{\mu}\|^2)\, z + \|\boldsymbol{\mu}\|^4}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right). \end{aligned} \tag{3.7}$$

When $\rho = \rho^*$, we observe that equation (3.7) is an odd function of z. Therefore, the function $g(\rho^*) = 0$, i.e. the integral of (3.7) is 0. To see that ρ^* is the only zero of g, we note that

$$g'(\rho) = \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\left(\boldsymbol{\mu}^T\boldsymbol{\xi}\right)^2 \exp(\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}{(1+\exp(\rho\boldsymbol{\mu}^T\boldsymbol{\xi}))^2}\right] > 0.$$

Here, $g'(\rho) = 0$ implies that $\boldsymbol{\mu}^T\boldsymbol{\xi} = 0$ a.s. which is not true. As a result, the function $g(\rho)$ is strictly decreasing with a zero at ρ^*. The result follows.

Hinge loss. We begin by noting that f is differentiable and it holds that

$$\nabla f(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\xi} 1_{\{\boldsymbol{\xi}^T\boldsymbol{\theta} \le 1\}}].$$

We next observe that for any $\boldsymbol{v}, \boldsymbol{\theta} \in \boldsymbol{R}^d$ such that $\boldsymbol{v}^T\boldsymbol{\theta} = 0$, it holds that

$$-\boldsymbol{v}^T \nabla f(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{v}^T\boldsymbol{\xi} 1_{\{\boldsymbol{\xi}^T\boldsymbol{\theta}\le 1\}}] = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{v}^T\boldsymbol{\xi}]\mathbb{E}_{\boldsymbol{\xi}}[1_{\{\boldsymbol{\xi}^T\boldsymbol{\theta}\le 1\}}] = \boldsymbol{v}^T\boldsymbol{\mu} \cdot \mathbb{E}_{\boldsymbol{\xi}}[1_{\{\boldsymbol{\xi}^T\boldsymbol{\theta}\le 1\}}]. \tag{3.8}$$

Here we used that $\boldsymbol{\xi}^T\boldsymbol{v}$ and $\boldsymbol{\xi}^T\boldsymbol{\theta}$ are independent random variables and the expectation of the product of two uncorrelated random variables is the product of the expectations. Now note that for any $\boldsymbol{\theta}$, the quantity $\mathbb{E}_{\boldsymbol{\xi}}[1_{\{\boldsymbol{\xi}^T\boldsymbol{\theta}\le 1\}}]$ is strictly positive. Therefore, if $\boldsymbol{v}^T\boldsymbol{\theta} = 0$ and $\nabla f(\boldsymbol{\theta}) = \mathbf{0}$ then, using (3.8), we obtain that $\boldsymbol{v}^T\boldsymbol{\mu} = 0$. Hence, we established that $\nabla f(\boldsymbol{\theta}) = \mathbf{0}$ implies $\boldsymbol{\theta} = \rho\boldsymbol{\mu}$ for some $\rho \in \boldsymbol{R}$. On the other hand, using (3.8) again, we have that $\nabla f(\rho\boldsymbol{\mu}) = 0$ if and only if $\boldsymbol{\mu}^T \nabla f(\rho\boldsymbol{\mu}) = 0$. To see the only if direction, suppose

$\boldsymbol{\mu}^T \nabla f(\rho\boldsymbol{\mu}) = 0$ and $\nabla f(\rho\boldsymbol{\mu}) \neq 0$. Then we have $\nabla f(\rho\boldsymbol{\mu}) = \boldsymbol{v}$ where the vector $\boldsymbol{v}$ is nonzero such that $\boldsymbol{v}^T\boldsymbol{\mu} = 0$. By (3.8), we deduce $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T \nabla f(\rho\boldsymbol{\mu}) = 0$ yielding a contradiction.

Next, consider the function

$$g(\rho) = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\mu}^T\boldsymbol{\xi} 1_{\{\rho\boldsymbol{\xi}^T\boldsymbol{\mu}\leq 1\}}]. \tag{3.9}$$

Observe that $g(\rho) = \boldsymbol{\mu}^T \nabla f(\rho\boldsymbol{\mu})$. Dominated Convergence Theorem yields that

$$\lim_{\rho\to+\infty} g(\rho) = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\mu}^T\boldsymbol{\xi} 1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}\leq 0\}}], \quad \lim_{\rho\to-\infty} g(\rho) = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\mu}^T\boldsymbol{\xi} 1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}\geq 0\}}].$$

It, therefore, holds that $\lim_{\rho\to+\infty} g(\rho) < 0$ and $\lim_{\rho\to-\infty} g(\rho) > 0$. Since $g(0) = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\mu}^T\boldsymbol{\xi}] > 0$, it remains to show that g is a strictly decreasing function. To this end, we note that for any fixed $\rho_1 < \rho_2$, it holds that

$$\boldsymbol{\mu}^T\boldsymbol{\xi}\left(1_{\{\rho_1\boldsymbol{\mu}^T\boldsymbol{\xi}\leq 1\}} - 1_{\{\rho_2\boldsymbol{\mu}^T\boldsymbol{\xi}\leq 1\}}\right) \geq 0 \quad \text{for any value of } \boldsymbol{\xi}. \tag{3.10}$$

Indeed, if $\boldsymbol{\mu}^T\boldsymbol{\xi} \geq 0$, then $\rho_1\boldsymbol{\mu}^T\boldsymbol{\xi} \leq \rho_2\boldsymbol{\mu}^T\boldsymbol{\xi}$; thus ensuring $1_{\{\rho_1\boldsymbol{\mu}^T\boldsymbol{\xi}\leq 1\}} \geq 1_{\{\rho_2\boldsymbol{\mu}^T\boldsymbol{\xi}\leq 1\}}$. The case $\boldsymbol{\mu}^T\boldsymbol{\xi} \leq 0$ follows similarly. We, therefore, conclude that $g(\rho_1) \geq g(\rho_2)$. Finally, note that $g(\rho_1) = g(\rho_2)$, implies that (3.10) holds with equality, almost surely. Clearly, this yields a contradiction. It remains to show (3.5). By (3.9), we have that $g'(\rho^*) = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\mu}^T\boldsymbol{\xi} 1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}\leq \frac{1}{\rho^*}\}}]$. Using (2.10) and (2.12), we obtain that

$$\Phi\left(\frac{1-\rho^*\|\boldsymbol{\mu}\|^2}{\rho^*\sigma\|\boldsymbol{\mu}\|}\right) \cdot \exp\left(\frac{1}{2}\cdot\left(\frac{1-\rho^*\|\boldsymbol{\mu}\|^2}{\rho^*\sigma\|\boldsymbol{\mu}\|}\right)^2\right) = \frac{1}{\sqrt{2\pi}}\cdot\frac{\sigma}{\|\mu\|}. \tag{3.11}$$

The result immediately follows.

□

We call a classifier, $\boldsymbol{\theta}^*$, *optimal* if it minimizes the probability of misclassification *i.e.*

$$\boldsymbol{\theta}^* \in \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{P}\left(\boldsymbol{\zeta}^T\boldsymbol{\theta} > 0 \,|\, \boldsymbol{\zeta} \sim \mathcal{P}_1\right), \tag{3.12}$$

We will use SGD with constant step-size to solve (3.4). Therefore, at each iteration, we query $\boldsymbol{\xi}_k \sim \mathcal{P}_*$ and updates the iterate based only on this sample as follows.

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \alpha\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}, 1). \tag{3.13}$$

Here $\alpha > 0$ is the algorithm's constant step-size. As explained in Chapter 2, with constant step-size, SGD only converges to a neighborhood of the minimizer. However, since the condition (3.12) is scale-invariant, for the task of binary classification one does not require convergence to a minimizer in order to obtain good classifiers and therefore constant step-size is favorable as it also eliminates the unnecessary effort for scheduling the learning rate.

3.2 Stopping criterion for SGD

Ordinarily in deterministic first-order optimization methods, one terminates when the norm of the gradient falls below a predefined tolerance. In the case of SGD for binary classification, this is unsuitable for two reasons. First, the true gradient is generally inaccessible to the algorithm or it is computationally expensive to generate even a sufficient approximation of the gradient. Second, even if the computations were possible, an 'optimal' classifier $\boldsymbol{\theta}$ for the classification task is not necessarily the minimizer of the loss function since the loss function is merely a surrogate for correct classification of the data. Note that, several works [16, 25, 24, 60, 61] have suggested an alternative for the stochastic setting– terminate when $\mathbb{P}(f(\boldsymbol{\theta}) - \min\ f \leq \varepsilon) \geq 1 - p$ for some chosen small $\varepsilon > 0$ and probability p.

For homogeneous linear classifiers applied to the hinge loss function, it has been shown ([57]) that the homotopic sub-gradient method converges to a maximal margin solution on linearly separable data. In ([58]), SGD applied to the logistic loss on linearly separable data will produce a sequence of $\boldsymbol{\theta}_k$ that diverge to infinity, but when normalized also converge to the L_2-max margin solution. Little is known about the behavior of constant step-size SGD when the linear separability assumption on the data is removed (see, *e.g.*, [90]). The assumption of zero-noise in our context would mean that $\mathcal{P}_0$, $\mathcal{P}_1$ each reduce to a single point, a trivial example of separable data. Since there is often noise in the sample procedure, the data *may not necessarily be linearly separable*. Understanding the behavior of SGD in the presence of noise is, therefore, important.

3.2.1 Stopping criterion

Even though the binary classifier is scale-free, the logistic and hinge regression loss is not. It transitions from flat to unit-slope when $\boldsymbol{\xi}^T\boldsymbol{\theta} = O(1)$. This suggests that when $\boldsymbol{\theta}$ reaches this region, a classification has been made. Motivated by this, we propose the following termination test: Sample $\hat{\boldsymbol{\xi}}_k \sim \mathcal{P}_*$ and

$$\text{Terminate when } \ \hat{\boldsymbol{\xi}}_k^T\boldsymbol{\theta}_k \geq 1. \tag{3.14}$$

However, the termination test (3.14) requires an additional sample and an additional inner product per iteration and, as such, imposes a small additional cost. To reduce this cost, in all our numerical experiments (Sec. 3.4), we use the test 3.1 which imposes no computational overhead as SGD already computes $\boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k$. We rewrite the termination criterion (3.1) below.

$$\text{Terminate when } \boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k \geq 1, \tag{3.15}$$

After testing both (3.14) and (3.15), we found that compared to the variation between successive randomized trials, their behaviors in numerical experiments were indistinguishable.

Notice that in support vector machine (SVM) theory [79], the scaling of the optimizing classifier is constrained so that the margin between classes is $O(1)$. Algorithm 8 describes the termination criteria (3.14) as applied with the update rule governed by SGD.

Algorithm 8: SGD with termination test

initialize: $\boldsymbol{\theta}_0 \in \boldsymbol{R}^d$, $\alpha > 0$, $\hat{\boldsymbol{\xi}}_0 \sim \mathcal{P}_*$, $k = 0$
while $\hat{\boldsymbol{\xi}}_k^T \boldsymbol{\theta}_k < 1$
Pick data point $\boldsymbol{\xi}_{k+1} \sim \mathcal{P}_*$.
Compute $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\xi}_{\boldsymbol{k+1}}^T \boldsymbol{\theta}_k, 1)$
Update $\boldsymbol{\theta}$ by setting

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\xi}_{\boldsymbol{k+1}}^T \boldsymbol{\theta}_k, 1) \tag{3.16}$$

Sample $\hat{\boldsymbol{\xi}}_{k+1} \sim \mathcal{P}_*$
$k \leftarrow k + 1$
end

Assumption 1. *[The distribution $\mathcal{P}_*$ is Gaussian]* Our theoretical analysis makes a further assumption on the distribution $\mathcal{P}_*$. For the rest of this section and Sec. 3.3, $\mathcal{P}_0 = N(\boldsymbol{\mu}_0, \sigma^2 I_d)$, $\mathcal{P}_1 = N(\boldsymbol{\mu}_1, \sigma^2 I_d)$, and therefore $\mathcal{P}_* = N(\boldsymbol{\mu}, \sigma^2 I_d)$, a Gaussian with unknown mean $\boldsymbol{\mu}$ $(= \boldsymbol{\mu}_1 = -\boldsymbol{\mu}_0)$ and variance $\sigma^2 I_d$. This assumption allows for non-separable data provided $\sigma > 0$.

Using Lemma 6, we give an exact characterization of the set of optimal classifiers (3.12) under Assumption 1. Note that (3.12) is rewritten as follows in terms of distribution $\mathcal{P}_*$.

$$\boldsymbol{\theta}^* \in \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{P}_{\boldsymbol{\xi} \sim \mathcal{P}_*} \left(\boldsymbol{\xi}^T \boldsymbol{\theta} > 0 \right), \tag{3.17}$$

Lemma 7. *Under Assumption 1, the set of optimal classifier defined in (3.17) equals to* $\{\lambda \boldsymbol{\theta}^* : \lambda > 0\}$.

Proof. Observe that the following simple fact holds.

$$\mathbb{P}_{\hat{\boldsymbol{\xi}}} \left(\hat{\boldsymbol{\xi}}^T \boldsymbol{\theta} \geq t \right) = \Phi^c \left(\frac{\boldsymbol{\mu}^T \boldsymbol{\theta} - t}{\sigma \|\boldsymbol{\theta}\|} \right), \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{R}^d, t \in \boldsymbol{R} \text{ and } \hat{\boldsymbol{\xi}} \sim N(\boldsymbol{\mu}, \sigma^2 I_d). \tag{3.18}$$

Therefore we have that $\mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\xi}^T\boldsymbol{\theta} > 0) = \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma} \cdot \cos(w_{\boldsymbol{\theta}})\right)$ where $\boldsymbol{\xi} \sim N(\boldsymbol{\mu}, \sigma^2 I_d)$ and $w_{\boldsymbol{\theta}}$ denotes the angle between the two vectors $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$. On the other hand a classifier $\boldsymbol{\theta}$ is optimal if and only if $\boldsymbol{\theta} = \rho\boldsymbol{\mu}$ for some $\rho > 0$, i.e. $\cos(w_{\boldsymbol{\theta}}) = 0$. The proof is complete after noting that Φ is an increasing function. □

3.3 Analysis of stopping criterion

In this section, we present our analysis of the stopping criterion (3.14) proposed in Section 3.2. Here we introduce the first iteration at which the stopping criterion is satisfied, denoted by the random variable

$$T := \inf\left\{k > 0 : \hat{\boldsymbol{\xi}}_k^T \boldsymbol{\theta}_k \geq 1\right\}. \tag{3.19}$$

By viewing the stopping criterion through the lens of stopping times, we are able to utilize probability theory to analyze the classifier at termination $\boldsymbol{\theta}_T$. Throughout this section, we work with the following filtration.

$$\mathcal{F}_0 = \sigma(\boldsymbol{\theta}_0) \quad \text{and} \quad \mathcal{F}_k := \sigma(\boldsymbol{\theta}_0, \hat{\boldsymbol{\xi}}_1, \boldsymbol{\xi}_1, \hat{\boldsymbol{\xi}}_2, \boldsymbol{\xi}_2, \ldots, \hat{\boldsymbol{\xi}}_k, \boldsymbol{\xi}_k), \quad \text{for all } k \geq 1 \tag{3.20}$$

Clearly, the random variable $\boldsymbol{\theta}_k$ is $\mathcal{F}_k$-measurable. Our theoretical results are structured as follows.

First, we show that SGD with our proposed termination test indeed stops after a finite number of iterations. To do so, we provide a bound on $\mathbb{E}[T]$, *i.e.* the expected number of iterations before termination. Yet, despite this guarantee, the resulting classifier at termination need not be optimal. Hence, our second result establishes that both $\boldsymbol{\theta}_T$ and $\boldsymbol{\theta}^*$ point in approximately the same direction; thereby ensuring that the classifier at termination, $\boldsymbol{\theta}_T$, is nearly optimal. We remark the worst-case bounds established throughout these sections are conservative; we observe in our experiments that the termination test stops sooner while also yielding good classification properties for Gaussian and non-Gaussian data sets.

To bound $\mathbb{E}[T]$, we identify subsets of $\boldsymbol{R}^d$ for which when an iterate enters the set, termination (*i.e.* (3.14)) is *highly likely* to succeed. Such sets C, we call *target sets*. Precisely, for any $\boldsymbol{\theta} \in C$ and $\hat{\boldsymbol{\xi}} \sim N(\boldsymbol{\mu}, \sigma^2 I_d)$, the probability of terminating is at least $\delta > 0$,

$$\exists\, \delta > 0 \text{ such that } \mathbb{P}_{\hat{\boldsymbol{\xi}}}\left(\hat{\boldsymbol{\xi}}^T\boldsymbol{\theta} \geq 1\right) \geq \delta. \tag{3.21}$$

We guarantee the iterates generated by SGD enter the target set by way of a *drift function*, $V : \boldsymbol{R}^d \to [0, +\infty)$. A drift function, on average, decreases each time the iterate fails to live in the target set. In other words, conditioned on the past iterates the following holds

$$(\mathbb{E}[V(\boldsymbol{\theta}_k)|\mathcal{F}_{k-1}] - V(\boldsymbol{\theta}_{k-1}))1_{\{\boldsymbol{\theta}_{k-1} \notin C\}} \leq -b1_{\{\boldsymbol{\theta}_{k-1} \notin C\}} \tag{3.22}$$

for the target set C and some positive constant b. Loosely speaking, the iterates in expectation *drift* towards the target set. Target sets and drift functions in the context of drift analysis are well-studied in stochastic processes [56].

A natural choice for the target set is a neighborhood of the unique optimum solution of (3.4), $\boldsymbol{\theta}^*$, with the drift function $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$. Indeed, it is known the iterates of SGD converge to a neighborhood of $\boldsymbol{\theta}^*$ ([70]). However, an iterate may be nearly optimal well before it enters this neighborhood. In fact when $\sigma \ll \|\boldsymbol{\mu}\|$, we identify a target set where satisfying the stopping criterion occurs at least half the time and does not require the iterate to be near $\boldsymbol{\theta}^*$. We summarize below our target set and drift function.

1. Under the assumption $\sigma \leq c\|\boldsymbol{\mu}\|$ for some numerical constant c, which we call the *Low Variance Regime*, we define the target set to be

$$C = \{\boldsymbol{\theta} : \boldsymbol{\mu}^T\boldsymbol{\theta} \geq 1\}, \tag{3.23}$$

   and the drift function by

$$V(\boldsymbol{\theta}) = \left(M - \boldsymbol{\mu}^T\boldsymbol{\theta}\right)^2, \tag{3.24}$$

   for some constant M, to be determined later.

2. Under the assumption $c\|\boldsymbol{\mu}\| \leq \sigma$ where the constant c is the same as in 1 above, which we call the *High Variance Regime*, we define the target set to be

$$C = \{\boldsymbol{\theta} : |\rho\sigma^2 - 1| < 1 \text{ and } \sigma\|\tilde{\boldsymbol{\theta}}\| \leq c'\}, \tag{3.25}$$

   for some numerical constant c'. Here, we orthogonally decompose $\boldsymbol{\theta} = \rho\boldsymbol{\mu} + \tilde{\boldsymbol{\theta}}$ with $\boldsymbol{\mu}^T\tilde{\boldsymbol{\theta}} = 0$. We use the following drift function

$$V(\boldsymbol{\theta}) = \frac{1}{2\alpha}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2. \tag{3.26}$$

In Section 3.3.1 (resp. Section 3.3.2) we show that the pairs (C, V) defined in (3.23) and (3.24) (resp. (3.25) and (3.26)) satisfies the drift equation (3.22) for any step-size α (resp. for any sufficiently small step-size α).

As mentioned above, the target set C attracts the iterates generated by SGD. Each time an iterate enters C, the stopping criterion holds with probability at least $\delta > 0$. Provided the iterates enters the set C an infinite number of times, then after waiting a geometrically distributed many iterations, we expect the following condition to hold:

$$\hat{\boldsymbol{\xi}}_k^T \boldsymbol{\theta}_k \geq 1 \text{ and } \boldsymbol{\theta}_k \in C. \tag{3.27}$$

The SGD algorithm does not know the value of $\boldsymbol{\theta}^*$; therefore at each iteration, it cannot check whether the condition (3.27) occurs. Nevertheless, we are able to compute a bound on the average waiting time until (3.27) holds and the first time (3.27) holds is always an upper bound on T, our stopping criterion. This is summarized in Lemma 8. Precisely, if we denote by

$$T_C := \inf\{k > 0 : \hat{\boldsymbol{\xi}}_k^T \boldsymbol{\theta}_k \geq 1 \text{ and } \boldsymbol{\theta}_k \in C\}, \tag{3.28}$$

then $T \leq T_C$, thus yielding $\mathbb{E}[T] \leq \mathbb{E}[T_C]$. We bound $\mathbb{E}[T_C]$ by way of stopping times τ_m defined as the m^{th} time the iterates of SGD enters C. Formally for any sequence $\{\boldsymbol{\theta}_k\}_{k=0}^{\infty}$ generated by SGD starting at $\boldsymbol{\theta}_0 = \mathbf{0}$, we set

$$\tau_1 := \inf\{k > 0 : \boldsymbol{\theta}_k \in C\} \tag{3.29}$$

and inductively, for $m \geq 2$,

$$\tau_m := \inf\{k > \tau_{m-1} : \boldsymbol{\theta}_k \in C\}. \tag{3.30}$$

The following lemma formalizes the discussion above.

Lemma 8. *Let $\{\boldsymbol{\theta}_k\}_{k=0}^{\infty}$ be a sequence generated by SGD such that $\boldsymbol{\theta}_0 = \mathbf{0}$ and suppose that $\mathbb{E}[\tau_m] < +\infty$ for all $m \geq 1$. Then the following holds*

$$\mathbb{E}[T] \leq \mathbb{E}[T_C] \leq \sum_{m=1}^{\infty} \mathbb{E}[\tau_m](1-\delta)^{m-1}, \tag{3.31}$$

where δ satisfies (3.21).

Proof. We first show that

$$\mathbb{E}\left[1_{\{T_C \geq \tau_m\}}\right] \leq (1-\delta)^{m-1}. \tag{3.32}$$

Define the σ-algebra $\mathcal{F}' = \sigma(\boldsymbol{\theta}_0, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots)$. From the independence between $\sigma(\hat{\boldsymbol{\xi}}_k)$'s and $\mathcal{F}'$ and also $\tau_i < +\infty$ a.s. for all $i \geq 1$, the following is obtained:

$$\begin{aligned}
\mathbb{E}\left[1_{\{T_C \geq \tau_m\}} | \mathcal{F}'\right] &= \mathbb{E}\left[1_{\{\hat{\boldsymbol{\xi}}_{\tau_1}^T \boldsymbol{\theta}_{\tau_1} < 1\}} \cdots 1_{\{\hat{\boldsymbol{\xi}}_{\tau_{m-1}}^T \boldsymbol{\theta}_{\tau_{m-1}} < 1\}} | \mathcal{F}'\right] \\
&= \prod_{i=1}^{m-1} \mathbb{E}\left[1_{\{\hat{\boldsymbol{\xi}}_{\tau_i}^T \boldsymbol{\theta}_{\tau_i} < 1\}} | \mathcal{F}'\right] \\
&\leq (1-\delta)^{m-1}.
\end{aligned}$$

By taking expectations, we conclude (3.32) holds. Now since $\mathbb{E}[1_{\{T_C=+\infty\}}] \leq \mathbb{E}[1_{\{T_C\geq\tau_m\}}]$ for all $m \geq 1$, it follows from (3.32) that $T_C < \infty$ a.s. We next observe that

$$\begin{aligned}
\mathbb{E}\left[T_C 1_{\{T_C=\tau_m\}}|\mathcal{F}'\right] &= \mathbb{E}\left[\tau_m 1_{\{T_C=\tau_m\}}|\mathcal{F}'\right] \\
&\leq \tau_m \mathbb{E}\left[1_{\{\hat{\boldsymbol{\xi}}_{\tau_1}^T \boldsymbol{\theta}_{\tau_1}<1\}} \cdots 1_{\{\hat{\boldsymbol{\xi}}_{\tau_{m-1}}^T \boldsymbol{\theta}_{\tau_{m-1}}<1\}}|\mathcal{F}'\right] \\
&= \tau_m \prod_{i=1}^{m-1} \mathbb{E}\left[1_{\{\hat{\boldsymbol{\xi}}_{\tau_i}^T \boldsymbol{\theta}_{\tau_i}<1\}}|\mathcal{F}'\right] \\
&\leq \tau_m (1-\delta)^{m-1}.
\end{aligned}$$

Taking expectations yields $\mathbb{E}\left[T_C 1_{\{T_C=\tau_m\}}\right] \leq \mathbb{E}\left[\tau_m\right](1-\delta)^{m-1}$ for all $m \geq 1$. Now since $T_C < \infty$ a.s. we get $1 = \sum_{m=1}^{+\infty} 1_{\{T_C=\tau_m\}}$ a.s. This yields that

$$\mathbb{E}[T] \leq \mathbb{E}[T_C] = \sum_{m=1}^{\infty} \mathbb{E}[T_C 1_{\{T_C=\tau_m\}}] \leq \sum_{m=1}^{\infty} \mathbb{E}[\tau_m](1-\delta)^{m-1}.$$

The proof is complete. □

Now, in view of Lemma 8, it suffices to bound $\mathbb{E}[\tau_m]$ by a sequence which can not grow too fast in m. Indeed, we show that (3.22) implies the following

$$\mathbb{E}[\tau_m] = \mathcal{O}(m). \tag{3.33}$$

Theorem 6. *(Low Regime) Let $\{\boldsymbol{\theta}_k\}_{k=0}^{\infty}$ be a sequence generated by Algorithm 8 such that $\boldsymbol{\theta}_0 = \mathbf{0}$. There exists positive constants c, b and M such that provided $\sigma \leq c\|\boldsymbol{\mu}\|$ the following holds.*

$$\mathbb{E}[T] \leq 2 + \frac{2M^2}{b} \cdot \left(\Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3}{\|\boldsymbol{\mu}\|} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right) + 1\right). \tag{3.34}$$

Here the constants c, b and M are defined as follows:

1. *For the logistic loss,*

$$c = 0.33, \quad b = \alpha\|\boldsymbol{\mu}\|^2, \quad \text{and } M = 501 + 640\alpha\|\boldsymbol{\mu}\|^2. \tag{3.35}$$

2. *For the hinge loss,*

$$c = 1.25, \quad b = \alpha\|\boldsymbol{\mu}\|^2, \quad \text{and } M = 501 + 782\alpha\|\boldsymbol{\mu}\|^2. \tag{3.36}$$

Therefore, on relatively separable data (*i.e.* in the low variance regime), the expected waiting time before termination exponentially decreases as the data becomes more separable (*i.e.* $\sigma \to 0$). We prove Theorem 6 in Section 3.3.3. The next theorem shows that the expected value of the stopping time is finite provided that the $\sigma > c\|\boldsymbol{\mu}\|$ and the step-size is small enough.

Theorem 7. *(High Regime) Suppose that $\sigma > c\|\boldsymbol{\mu}\|$ where c is defined in (3.35) and (3.36). Then there exists a universal positive constant A such that if the step-size α satisfies*

$$\alpha \leq A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2(\|\boldsymbol{\mu}\|^2 + d\sigma^2)}, \tag{3.37}$$

then it holds that $\mathbb{E}[T] < +\infty$. In particular, the termination criterion occurs almost surely.

It remains to determine whether the classifier at termination $\boldsymbol{\theta}_T$, has desirable accuracy. The scale-invariance of optimal classifiers means a classifier yields a lower probability of misclassification the closer its direction aligns with any optimal classifier. In view of this, it suffices to bound the absolute value of the inner product of any unit vector that is perpendicular to $\boldsymbol{\theta}^*$, $\boldsymbol{v}$ with $\boldsymbol{\theta}_T$. The following theorem establishes a bound on $\mathbb{E}[|\boldsymbol{v}^T\boldsymbol{\theta}_T|]$.

Theorem 8. *Let $\boldsymbol{\theta}_0 = \mathbf{0}$. Fix any unit vector $\boldsymbol{v} \in \boldsymbol{R}^d$ such that $\boldsymbol{v}^T\boldsymbol{\theta}^* = 0$. Then the following estimate holds*

$$\mathbb{E}[|\boldsymbol{v}^T\boldsymbol{\theta}_T|] \leq \sigma\alpha\sqrt{\frac{2}{\pi}}\mathbb{E}[T]. \tag{3.38}$$

In the low variance regime by combining Theorem 6 and 8 for a fixed step-size α it holds that $\mathbb{E}[|\boldsymbol{v}^T\boldsymbol{\theta}|] \leq \mathcal{O}(\sigma)$. Thus, the more separable the data set is, the more accurate the classifier $\boldsymbol{\theta}_T$ is on average. In the high variance regime, Theorem 7 yields a very loose bound. Yet despite this, our numerical result in Section 3.4 show promising accuracy of (3.14) in this case as well. We conjecture that the inequality can be significantly strengthened.

3.3.1 Low regime, proof of Theorem 6

In this section, we investigate the low variance regime. We consider the target set C and function V defined in (3.23) and (3.24) respectively, *i.e.*

$$C = \{\boldsymbol{\theta} : \boldsymbol{\mu}^T\boldsymbol{\theta} \geq 1\}, \quad V(\boldsymbol{\theta}) = \left(M - \boldsymbol{\mu}^T\boldsymbol{\theta}\right)^2, \tag{3.39}$$

where M is a constant to be determined. Next lemma shows that the drift equation (3.22) holds for the pair (C, V).

Lemma 9 (Drift equation)**.** *Consider the SGD algorithm and let the set C and the function V be as in (3.39). Define the constants c, b, M as in (3.35) and (3.36). Then provided that $\sigma \leq c\|\boldsymbol{\mu}\|$, the function V is a drift function with respect to the set C and it satisfies the drift equation (3.22) with the constant b.*

Proof. For simplicity we write $\mathcal{F}_{-1} := \sigma(\{\boldsymbol{\theta}_0 = \boldsymbol{\theta}\})$. Fix $k \geq 1$ and write $\boldsymbol{\xi}_k = \boldsymbol{\mu} + \sigma\boldsymbol{\psi}_k$ with $\boldsymbol{\psi}_k \sim N(0, I_d)$. Denote $\psi_k := \frac{\boldsymbol{\mu}^T\boldsymbol{\psi}_k}{\|\boldsymbol{\mu}\|}$, thus $\psi_k \sim N(0,1)$. In order to show that the function V satisfies the drift equation (3.22), it suffices to assume $\boldsymbol{\theta}_{k-1} \notin C$; in particular, this means $\boldsymbol{\theta}_{k-1}^T\boldsymbol{\mu} < 1$.

Logistic loss. By expanding out the term using the update formula, we get the following

$$V(\boldsymbol{\theta}_k) = V(\boldsymbol{\theta}_{k-1}) - \frac{2\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_k(M - \boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1})}{1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})} + \frac{\alpha^2(\boldsymbol{\mu}^T\boldsymbol{\xi}_k)^2}{(1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}))^2}. \tag{3.40}$$

We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}_k}\left[\frac{\boldsymbol{\mu}^T\boldsymbol{\xi}_k}{1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})}|\mathcal{F}_{k-1}\right] \\
&\quad= \|\boldsymbol{\mu}\|^2\mathbb{E}_{\boldsymbol{\xi}_k}\left[\frac{1}{1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})}|\mathcal{F}_{k-1}\right] + \sigma\|\boldsymbol{\mu}\|\mathbb{E}_{\boldsymbol{\xi}_k,\psi_k}\left[\frac{\psi_k}{1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})}|\mathcal{F}_{k-1}\right] \\
&\quad\geq \|\boldsymbol{\mu}\|^2\mathbb{E}_{\boldsymbol{\xi}_k}\left[\frac{1}{1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})}|\mathcal{F}_{k-1}\right] + \sigma\|\boldsymbol{\mu}\|\mathbb{E}_{\psi_k}\left[\psi_k 1_{\{\psi_k<0\}}\right] \\
&\quad= \|\boldsymbol{\mu}\|^2\mathbb{E}_{\boldsymbol{\xi}_k}\left[\frac{1}{1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})}\left(1_{\{\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1}\geq\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\}} + 1_{\{\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1}<\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\}}\right)|\mathcal{F}_{k-1}\right] - \sigma\|\boldsymbol{\mu}\|\sqrt{\frac{1}{2\pi}} \\
&\quad\geq \frac{\|\boldsymbol{\mu}\|^2}{1+\exp(\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1})}\mathbb{E}_{\boldsymbol{\xi}_k}\left[1_{\{\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1}\geq\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\}}|\mathcal{F}_{k-1}\right] - \sigma\|\boldsymbol{\mu}\|\sqrt{\frac{1}{2\pi}} \\
&\quad\geq \frac{\|\boldsymbol{\mu}\|^2}{2(1+e)} - \sigma\|\boldsymbol{\mu}\|\sqrt{\frac{1}{2\pi}} \\
&\quad\geq 0.001\|\boldsymbol{\mu}\|^2.
\end{aligned}$$

Here the first inequality follows from $\mathbb{E}[X] \geq \mathbb{E}[X1_{\{X<0\}}]$ and $1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}) \geq 1$, the second equation from (2.13), and the second to last from the observation that for any X normally distributed, $\mathbb{P}(\mathbb{E}[X] \geq X) = 1/2$ and $\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1} \sim N(\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1}, \sigma^2\|\boldsymbol{\theta}_{k-1}\|^2)$ and $\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1} < 1$. The last inequality uses the assumption $\sigma \leq 0.33\|\boldsymbol{\mu}\|$. By taking the conditional expectations of (3.40) combined with the above sequence of inequalities, we

deduce the following bound

$$\begin{aligned}
&\mathbb{E}\left[V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k-1})|\mathcal{F}_{k-1}\right] \\
&\quad = \mathbb{E}_{\boldsymbol{\xi}_k}\left[-\frac{2\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_k(M - \boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1})}{1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})}|\mathcal{F}_{k-1}\right] + \mathbb{E}_{\boldsymbol{\xi}_k}\left[\frac{\alpha^2(\boldsymbol{\mu}^T\boldsymbol{\xi}_k)^2}{(1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}))^2}|\mathcal{F}_{k-1}\right] \\
&\quad \leq -0.002(M-1)\alpha\|\boldsymbol{\mu}\|^2 + \alpha^2\|\boldsymbol{\mu}\|^2\left(\|\boldsymbol{\mu}\|^2 + \sigma^2\right) \\
&\quad = \alpha\|\boldsymbol{\mu}\|^2\left[-0.002(M-1) + \alpha\left(\|\boldsymbol{\mu}\|^2 + \sigma^2\right)\right].
\end{aligned}$$

Here the first inequality follows from $\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1} < 1$ and by upper bounding $\frac{(\boldsymbol{\mu}^T\boldsymbol{\xi}_k)^2}{(1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}))^2}$ with $(\boldsymbol{\mu}^T\boldsymbol{\xi}_k)^2$ and then applying (2.13). A quick computation after plugging in the value of M and the bound $\sigma \leq 0.33\|\boldsymbol{\mu}\|$ from (3.35) yields the drift equation (3.22) with $b = \alpha\|\boldsymbol{\mu}\|^2$.

Hinge loss. By expanding out the term using the update formula, we get the following

$$V(\boldsymbol{\theta}_k) = V(\boldsymbol{\theta}_{k-1}) - 2\alpha(M - \boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1})\boldsymbol{\mu}^T\boldsymbol{\xi}_k 1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}} + \alpha^2(\boldsymbol{\mu}^T\boldsymbol{\xi}_k)^2 1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}. \tag{3.41}$$

We have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}_k}[1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}\boldsymbol{\mu}^T\boldsymbol{\xi}_k|\mathcal{F}_{k-1}] &= \|\boldsymbol{\mu}\|^2\mathbb{E}_{\boldsymbol{\xi}_k}[1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}|\mathcal{F}_{k-1}] + \sigma\|\boldsymbol{\mu}\|\mathbb{E}_{\boldsymbol{\xi}_k,\psi_k}[1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}\psi_k|\mathcal{F}_{k-1}] \\
&\geq \frac{1}{2}\|\boldsymbol{\mu}\|^2 + \sigma\|\boldsymbol{\mu}\|\mathbb{E}_{\psi_k}[\psi_k 1_{\{\psi_k<0\}}] \\
&= \frac{1}{2}\|\boldsymbol{\mu}\|^2 - \sigma\|\boldsymbol{\mu}\|\sqrt{\frac{1}{2\pi}} \\
&\geq 0.001\|\boldsymbol{\mu}\|^2.
\end{aligned}$$

Here the first inequality follows from $1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1}\}} \leq 1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}$ and $\mathbb{E}_{\boldsymbol{\xi}_k}[1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1}\}}] = \frac{1}{2}$, and the second from (2.13). The last inequality uses the assumption $\sigma \leq 1.25\|\boldsymbol{\mu}\|$. By taking conditional expectations of (3.41) combined with the above sequence of inequalities, we deduce the bound

$$\begin{aligned}
\mathbb{E}[V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k-1})|\mathcal{F}_{k-1}] &= \mathbb{E}_{\boldsymbol{\xi}_k}\left[-2\alpha(M - \boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1})1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}|\mathcal{F}_{k-1}\right] \\
&\quad + \mathbb{E}_{\boldsymbol{\xi}_k}\left[\alpha^2(\boldsymbol{\mu}^T\boldsymbol{\xi}_k)^2 1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}|\mathcal{F}_{k-1}\right] \\
&\quad \leq \alpha\|\boldsymbol{\mu}\|^2\left[-0.002(M-1) + \alpha\left(\|\boldsymbol{\mu}\|^2 + \sigma^2\right)\right].
\end{aligned}$$

A quick computation after plugging in the value of M and the bound $\sigma \leq 1.25\|\boldsymbol{\mu}\|$ yields the desired result. □

Recall, the stopping times τ_m denote the m^{th} time that the SGD iterates enter the target set C. We show that $\mathbb{E}[\tau_m] = \mathcal{O}(m)$. To do so, we begin by stating a lemma that gives a bound on the stopping time $\tilde{\tau}_1$ starting from any $\boldsymbol{\theta}_0$. In other words, for an arbitrary starting $\boldsymbol{\theta}_0$, we define

$$\tilde{\tau}_1 := \inf\{k > 0 : \boldsymbol{\theta}_k \in C\}.$$

We now establish upper bounds on $\mathbb{E}[\tau_m]$ for $m \geq 1$ in the following proposition.

Proposition 2. *(Bound on $\mathbb{E}[\tau_m]$) Let $\boldsymbol{\theta}_0 = \mathbf{0}$ and assume the notation and assumptions of Lemma 9 hold. The following is true for all $m \geq 1$*

$$\mathbb{E}[\tau_m] \leq (m-1)\left(1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3 M^2}{\|\boldsymbol{\mu}\| b} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right)\right) + \frac{M^2}{b}. \quad (3.42)$$

Proof. First, the result for $m = 1$ follows immediately by combining Lemma 9 and Proposition 1 with $\boldsymbol{\theta}_0 = \mathbf{0}$. We now assume that $\tau_{m-1} < \infty$ a.s. for some $m \geq 2$. Fix an integer $n \geq 1$. We decompose the space to yield the following bounds

$$\begin{aligned}
\mathbb{E}\left[(\tau_m - \tau_{m-1}) \wedge n | \mathcal{F}_{\tau_{m-1}+1}\right] &= \mathbb{E}\left[((\tau_m - \tau_{m-1}) \wedge n) | \mathcal{F}_{\tau_{m-1}+1}\right] 1_{\{\boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} \geq 1\}} \\
&\quad + \mathbb{E}\left[((\tau_m - \tau_{m-1}) \wedge n) | \mathcal{F}_{\tau_{m-1}+1}\right] 1_{\{\boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} < 1\}} \\
&= 1_{\{\boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} \geq 1\}} + \mathbb{E}\left[((\tau_m - \tau_{m-1}) \wedge n) | \mathcal{F}_{\tau_{m-1}+1}\right] 1_{\{\boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} < 1\}} \\
&= 1_{\{\boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} \geq 1\}} + \sum_{i=1}^{\infty} \mathbb{E}\left[(\tau_m - \tau_{m-1}) \wedge n | \mathcal{F}_{\tau_{m-1}+1}\right] 1_{\{i-1 < 1 - \boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} \leq i\}} \\
&= 1 + \sum_{i=1}^{\infty} \mathbb{E}\left[\tilde{\tau}_1 \wedge n | \boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\tau_{m-1}+1}\right] 1_{\{i-1 < 1 - \boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} \leq i\}}.
\end{aligned} \quad (3.43)$$

Here the first equality follows because $((\tau_m - \tau_{m-1}) \wedge n) 1_{\{\boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} \geq 1\}} = 1_{\{\boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} \geq 1\}}$ and the last equality by the strong Markov property. We consider the logistic and hinge loss case separately to show that the following is true

$$1_{\{i-1 < 1 - \boldsymbol{\mu}^T \boldsymbol{\theta}_{\tau_{m-1}+1} \leq i\}} \leq 1_{\{\boldsymbol{\mu}^T \boldsymbol{\xi}_{\tau_{m-1}+1} < \frac{1-i}{\alpha}\}}. \quad (3.44)$$

For clarity, in the next few inequalities, we write $1\{.\}$ instead of $1_{\{.\}}$. In case of logistic

loss, for each $i \geq 1$, we observe the bound

$$\begin{aligned}
1\{i-1 < 1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1} \leq i\} &\leq 1\{i-1 < 1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\} \\
&= 1\left\{i-1 < 1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}} - \frac{\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}}{1+\exp(\boldsymbol{\xi}_{\tau_{m-1}+1}^T\boldsymbol{\theta}_{\tau_{m-1}})}\right\} \\
&\leq 1\left\{i-1 < -\frac{\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}}{1+\exp(\boldsymbol{\xi}_{\tau_{m-1}+1}^T\boldsymbol{\theta}_{\tau_{m-1}})}\right\} \\
&\leq 1\left\{i-1 < -\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}\right\},
\end{aligned}$$

where the second inequality follows because $\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}} \geq 1$ and the last inequality because $-\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}$ is positive since $i-1 \geq 0$.

In case of hinge loss, for each $i \geq 1$, similar as above, we observe the bound

$$\begin{aligned}
1\left\{i-1 < 1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1} \leq i\right\} &\leq 1\left\{i-1 < 1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\right\} \\
&\leq 1\left\{i-1 < 1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}} - \alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}1_{\{\boldsymbol{\xi}_{\tau_{m-1}+1}^T\boldsymbol{\theta}_{\tau_{m-1}}\leq 1\}}\right\} \\
&\leq 1\left\{i-1 < -\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}1_{\{\boldsymbol{\xi}_{\tau_{m-1}+1}^T\boldsymbol{\theta}_{\tau_{m-1}}\leq 1\}}\right\} \\
&= 1\left\{i-1 < -\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}\right\}.
\end{aligned} \tag{3.45}$$

Therefore we have shown that (3.44) holds. Setting $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\tau_{m-1}+1}$, by Proposition 1 for each $i \geq 1$, we deduce

$$\begin{aligned}
\mathbb{E}\left[\tilde{\tau}_1 \wedge n | \boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\tau_{m-1}+1}\right] 1_{\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\leq i\}} &\leq \frac{(M-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1})^2}{b} 1_{\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\leq i\}} \\
&\leq \frac{(M+i-1)^2}{b} 1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}<\frac{1-i}{\alpha}\}}.
\end{aligned} \tag{3.46}$$

Finally we observe that

$$\begin{aligned}
\mathbb{E}\left[1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}<\frac{1-i}{\alpha}\}}\right] &= \mathbb{E}\left[\sum_{k=1}^{\infty} 1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{k+1}<\frac{1-i}{\alpha}\}} 1_{\{\tau_{m-1}=k\}}\right] \\
&= \sum_{k=1}^{\infty} \mathbb{E}\left[1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{k+1}<\frac{1-i}{\alpha}\}}\right] \mathbb{E}\left[1_{\{\tau_{m-1}=k\}}\right] \\
&= \Phi\left(\frac{\frac{1-i}{\alpha} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}\right) \sum_{k=1}^{\infty} \mathbb{E}\left[1_{\{\tau_{m-1}=k\}}\right] \\
&= \Phi\left(\frac{\frac{1-i}{\alpha} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}\right).
\end{aligned} \tag{3.47}$$

The second equality is by independence and the third equality because $\boldsymbol{\mu}^T\boldsymbol{\xi}_{k+1} \sim N(\|\boldsymbol{\mu}\|^2, \sigma^2\|\boldsymbol{\mu}\|^2)$. By combining (3.43), (3.46), and (3.47), we obtain the following

$$\begin{aligned}
\mathbb{E}\big[(\tau_m - \tau_{m-1}) \wedge n\big] &\leq 1 + \frac{M^2}{b} \cdot \Phi\left(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \sum_{i=2}^{\infty} \frac{(M+i-1)^2}{b} \cdot \Phi\left(\frac{\frac{1-i}{\alpha} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}\right) \\
&= 1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \sum_{i=2}^{\infty} \frac{(M+i-1)^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|^2 + \frac{i-1}{\alpha}}{\sigma\|\boldsymbol{\mu}\|}\right) \\
&\leq 1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma\|\boldsymbol{\mu}\|}{b\sqrt{2\pi}} \cdot \sum_{i=2}^{\infty} \frac{(M+i-1)^2}{\alpha\|\boldsymbol{\mu}\|^2 + i - 1} \cdot \exp\left(-\frac{1}{2}\left(\frac{\|\boldsymbol{\mu}\|^2 + \frac{i-1}{\alpha}}{\sigma\|\boldsymbol{\mu}\|}\right)^2\right),
\end{aligned} \tag{3.48}$$

where we used the inequality $\Phi^c(t) < \frac{1}{t\sqrt{2\pi}}\exp(-\frac{t^2}{2})$ for all $t > 0$. Next, note that $\frac{M+i-1}{\alpha\|\boldsymbol{\mu}\|^2+i-1} \leq \frac{M}{\alpha\|\boldsymbol{\mu}\|^2}$ holds for all $i \geq 2$. Using this we obtain the following bound

$$\begin{aligned}
&\sum_{i=2}^{\infty} \frac{(M+i-1)^2}{\alpha\|\boldsymbol{\mu}\|^2 + i - 1} \cdot \exp\left(-\frac{1}{2}\left(\frac{\|\boldsymbol{\mu}\|^2 + \frac{i-1}{\alpha}}{\sigma\|\boldsymbol{\mu}\|}\right)^2\right) \\
&\qquad \leq \frac{\sigma M^2}{\alpha\|\boldsymbol{\mu}\|^3} \cdot \sum_{i=2}^{\infty} \frac{\alpha\|\boldsymbol{\mu}\|^2 + i - 1}{\alpha\sigma\|\boldsymbol{\mu}\|} \cdot \exp\left(-\frac{1}{2}\left(\frac{\alpha\|\boldsymbol{\mu}\|^2 + i - 1}{\alpha\sigma\|\boldsymbol{\mu}\|}\right)^2\right) \\
&\qquad \leq \frac{\sigma M^2}{\alpha\|\boldsymbol{\mu}\|^3} \cdot \alpha\sigma\|\boldsymbol{\mu}\| \cdot \int_{\frac{\|\boldsymbol{\mu}\|}{\sigma}}^{+\infty} t\exp\left(-\frac{t^2}{2}\right) dt \\
&\qquad = \frac{\sigma^2 M^2}{\|\boldsymbol{\mu}\|^2} \cdot \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right).
\end{aligned} \tag{3.49}$$

Here we have used that $t \mapsto t\exp(-\frac{t^2}{2})$ is decreasing over $[1, +\infty)$. Combining (3.48) and (3.49), we obtain that

$$\mathbb{E}\left[(\tau_m - \tau_{m-1}) \wedge n\right] \leq 1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3 M^2}{\|\boldsymbol{\mu}\| b} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right). \tag{3.50}$$

Taking the limit as $n \to +\infty$, we observe that

$$\mathbb{E}[\tau_m] \leq 1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3 M^2}{\|\boldsymbol{\mu}\| b} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right) + \mathbb{E}[\tau_{m-1}].$$

We then iterate the above inequality yielding

$$\mathbb{E}[\tau_m] \leq (m-1)\left(1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3 M^2}{\|\boldsymbol{\mu}\| b} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right)\right) + \mathbb{E}[\tau_1].$$

The result follows by plugging in the bound from Proposition 1 for the base case $m = 1$. □

We are now ready to prove Theorem 6.

Proof of Theorem 6. In order to simplify the subsequent argument, we define the quantity,

$$M' := 1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3 M^2}{\|\boldsymbol{\mu}\| b} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right).$$

It is easy to see that

$$\mathbb{P}_{\hat{\boldsymbol{\xi}} \sim N(\boldsymbol{\mu}, \sigma^2 I_d)}\left(\hat{\boldsymbol{\xi}}^T \boldsymbol{\theta} \geq 1\right) \geq \frac{1}{2} \text{ for any } \boldsymbol{\theta} \in C.$$

Therefore $\delta = \frac{1}{2}$ satisfies (3.21). By Proposition 2 with Lemma 8, we conclude that

$$\mathbb{E}[T] \leq \mathbb{E}[T_C] = \sum_{m=1}^{\infty} \mathbb{E}[T_C 1_{\{T_C = \tau_m\}}] \leq \sum_{m=1}^{\infty} \frac{\mathbb{E}[\tau_m]}{2^{m-1}} \leq \sum_{m=1}^{\infty} \frac{(m-1)M' + \frac{M^2}{b}}{2^{m-1}} = 2M' + \frac{2M^2}{b}.$$

□

3.3.2 High regime, proof of Theorem 7

In this section, we consider the high variance regime. We consider the target set C and the function V defined in (3.25) and (3.26), respectively, *i.e.*

$$C := \left\{ \boldsymbol{\theta} : |\rho - \rho^*| < \tfrac{1}{2}\rho^* \text{ and } \sigma\|\tilde{\boldsymbol{\theta}}\| \leq c' \right\} \quad \text{and} \quad V(\boldsymbol{\theta}) := \frac{1}{2\alpha}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2, \tag{3.51}$$

where the minimizer $\boldsymbol{\theta}^* = \rho^*\boldsymbol{\mu}$ is defined in Lemma 6 and the constant c' is to be determined. We first aim to show that V is a drift function with respect to the set C under the high variance regime assumption, meaning $\sigma \geq c\|\boldsymbol{\mu}\|$. We next state a standard SGD convergence result applied to the logistic and hinge loss functions. We need the following technical lemma below.

Lemma 10. *Consider the following convex optimization problem*

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \tag{3.52}$$

Assume that $\boldsymbol{\theta}^*$ denotes the unique minimizer of f in (3.52) and the unbiased stochastic gradients $\boldsymbol{g_k}$ satisfy $\mathbb{E}\left[\|\boldsymbol{g_k}\|^2\right] \leq B$. Let $\boldsymbol{\theta}_0 \in \boldsymbol{R}^d$. The sequence $\{\boldsymbol{\theta}_k\}_{k=0}^{\infty}$ generated by SGD with fixed step-sizes α satisfies the following for all $k \geq 1$,

$$f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) \leq \frac{1}{2\alpha}\left(\|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\|^2 - \mathbb{E}\left[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2 |\mathcal{F}_{k-1}\right]\right) + \frac{\alpha}{2} \cdot B. \tag{3.53}$$

Proof. We begin by observing that

$$\boldsymbol{g_k} := \frac{1}{\alpha}\left(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_k\right)$$

and also

$$\mathbb{E}\left[\boldsymbol{g}_k|\mathcal{F}_{k-1}\right] = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_{k-1}).$$

By convexity of the function f, we have the following

$$\begin{aligned}
\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2 &= \|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\|^2 - 2\alpha \boldsymbol{g}_k^T(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*) + \alpha^2 \|\boldsymbol{g}_k\|^2 \\
&= \|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\|^2 - 2\alpha(\boldsymbol{g}_k - \mathbb{E}\left[\boldsymbol{g}_k|\mathcal{F}_{k-1}\right])^T(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*) - 2\alpha\mathbb{E}\left[\boldsymbol{g}_k|\mathcal{F}_{k-1}\right]^T(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*) \\
&\quad + \alpha^2 \|\boldsymbol{g}_k\|^2 \\
&\leq \|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\|^2 - 2\alpha(\boldsymbol{g}_k - \mathbb{E}\left[\boldsymbol{g}_k|\mathcal{F}_{k-1}\right])^T(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*) - 2\alpha(f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*)) \\
&\quad + \alpha^2 \|\boldsymbol{g}_k\|^2 .
\end{aligned}$$

By taking conditional expectations with respect to $\mathcal{F}_{k-1}$ and rearranging the above inequality, we obtain that

$$f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) \leq \frac{1}{2\alpha}\left(\|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\|^2 - \mathbb{E}\left[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2 \,|\mathcal{F}_{k-1}\right]\right) + \frac{\alpha}{2}\mathbb{E}\left[\|\boldsymbol{g_k}\|^2\right]. \tag{3.54}$$

The result follows immediately. □

By Lemma 10 for each $k \geq 1$, we deduce

$$\mathbb{E}[V(\boldsymbol{\theta}_k)|\mathcal{F}_{k-1}] - V(\boldsymbol{\theta}_{k-1}) \leq -\left(f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*)\right) + \frac{\alpha}{2}(\|\boldsymbol{\mu}\|^2 + d\sigma^2). \tag{3.55}$$

Therefore, in order to show that the pair (C, V) in (3.51) satisfies the drift equation (3.22), it suffices to lower bound the quantity $f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*)$ whenever $\boldsymbol{\theta}_{k-1} \notin C$. To do so, we orthogonally decompose $\boldsymbol{\theta}_{k-1} = \rho_{k-1}\boldsymbol{\mu} + \tilde{\boldsymbol{\theta}}_{k-1}$, *i.e.* $\boldsymbol{\mu}^T\tilde{\boldsymbol{\theta}}_{k-1} = 0$ and $\rho_{k-1} \in \boldsymbol{R}$ and write

$$f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) = \underbrace{f(\boldsymbol{\theta}_{k-1}) - f(\rho_{k-1}\boldsymbol{\mu})}_{(a)} + \underbrace{f(\rho_{k-1}\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*)}_{(b)}. \tag{3.56}$$

The assumption $\boldsymbol{\theta}_{k-1} \notin C$ yields that either $\sigma\|\tilde{\boldsymbol{\theta}}_{k-1}\| \geq c'$ or $|\rho_{k-1} - \rho^*| \geq \frac{1}{2}\rho^*$. In Lemma 11 (resp. 13), we show that (a) (resp. (b)) in (3.56) are both non-negative and they are lower bounded by some positive constant provided that $\sigma\|\tilde{\boldsymbol{\theta}}_{k-1}\| \geq c'$ and $|\rho_{k-1} - \rho^*| \leq \frac{1}{2}\rho^*$ (resp. $|\rho_{k-1} - \rho^*| \geq \frac{1}{2}\rho^*$).

Lemma 11. *(Lower bound for (a) in (3.56)) Fix $\boldsymbol{\theta} \in \boldsymbol{R}^d$ and orthogonally decompose $\boldsymbol{\theta} = \rho\boldsymbol{\mu} + \tilde{\boldsymbol{\theta}}$ where $\boldsymbol{\mu}^T\tilde{\boldsymbol{\theta}} = 0$ and $\rho \in \boldsymbol{R}$. Then the following are true*

1. $f(\boldsymbol{\theta}) - f(\rho\boldsymbol{\mu}) \geq 0$.
2. $f(\boldsymbol{\theta}) - f(\rho\boldsymbol{\mu}) \geq 1$ *provided that* $|\rho - \rho^*| \leq \frac{1}{2}\rho^*$, $\sigma\|\tilde{\boldsymbol{\theta}}\| \geq c'$ *and* $\sigma \geq c\|\boldsymbol{\mu}\|$ *where c is defined in (3.35) and (3.36). Here ρ^* is defined in Lemma 6 and the constant c' is defined by* 436 *and* $8 + 10\rho^*\sigma^2$ *for the logistic and hinge loss respectively.*

Proof. We consider the logistic and hinge loss separately.

Logistic loss. The two normal random variables, $\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi} \sim N(0, \sigma^2\|\tilde{\boldsymbol{\theta}}\|^2)$ and $\boldsymbol{\mu}^T\boldsymbol{\xi} \sim N(\|\boldsymbol{\mu}\|^2, \sigma^2\|\boldsymbol{\mu}\|^2)$, are independent by (2.11). Since we have

$$\mathbb{E}_{\boldsymbol{\xi}}[\log(\exp(-\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}))] = \mathbb{E}_{\boldsymbol{\xi}}[\log(\exp(\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}))] = 0,$$

it holds that

$$\begin{aligned} f(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})\right)\right] &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\exp(-\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi})\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\exp(\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi})+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\exp(-\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi})+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)\right], \end{aligned}$$

where the last equality is true because $\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi} \sim -\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}$. Therefore we obtain

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})\right)\right] \\ &\quad= \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\exp(\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi})+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\exp(-\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi})+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)\right] \\ &\quad= \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left((\exp(\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi})+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}))(\exp(-\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi})+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}))\right)\right] \\ &\quad= \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\exp(-\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})+\exp(\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})+\exp(-2\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)\right]. \end{aligned}$$

By the equality $\exp(\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi})+\exp(-\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}) = 2+4\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})$, we have

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})\right)\right] \\ &\quad= \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+2\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})+\exp(-2\rho\boldsymbol{\mu}^T\boldsymbol{\xi})+4\sinh^2(\tfrac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)\right]. \end{aligned}$$

Therefore, we have

$$\begin{aligned} 2\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\frac{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})}{1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right] &= 2\mathbb{E}_{\boldsymbol{\xi}}\left[\log(1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi}))\right] - \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)^2\right] \\ &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\frac{4\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}{(1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}))^2}\right)\right] \geq 0. \end{aligned} \tag{3.57}$$

Thereby, we showed that $f(\boldsymbol{\theta}) - f(\rho\boldsymbol{\mu}) \geq 0$. Now we establish the positive lower bound. First, we note the following

$$1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}) = 2\exp(-\tfrac{\rho\boldsymbol{\mu}^T\boldsymbol{\xi}}{2})\cosh(\tfrac{\rho\boldsymbol{\mu}^T\boldsymbol{\xi}}{2}).$$

Fix a constant $r > 0$ and consider the set $\{\boldsymbol{\xi} : |\boldsymbol{\theta}^T\boldsymbol{\xi}| > r\}$. Applying the inequality

$x^2+y^2 \geq 2|xy|$ and (3.57), we obtain that

$$
\begin{aligned}
2\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\frac{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})}{1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right] &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\frac{4\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}{(1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}))^2}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\frac{\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})}{\cosh^2(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right] \\
&\geq \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\frac{\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})}{\cosh^2(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\cdot 1_{\{\boldsymbol{\xi}:|\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}|\geq r\}}\right] \qquad (3.58) \\
&\geq \mathbb{E}_{\boldsymbol{\xi}}\left[\left(\log 2+\log\left(\frac{|\sinh(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})|}{\cosh(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right)\cdot 1_{\{\boldsymbol{\xi}:|\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}|\geq r\}}\right].
\end{aligned}
$$

Here (3.58) follows from $\log\left(1+\frac{\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})}{\cosh^2(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)$ is always positive. From (2.10), we have

$$
\boldsymbol{\mu}^T\boldsymbol{\xi} \sim N(\|\boldsymbol{\mu}\|^2, \sigma^2\|\boldsymbol{\mu}\|^2) \text{ and } \tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi} \sim N(0, \sigma^2\|\tilde{\boldsymbol{\theta}}\|^2).
$$

Therefore, $\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi} = \sigma\|\tilde{\boldsymbol{\theta}}\|\psi$ where $\psi \sim N(0,1)$. Moreover, a simple computation shows that

$$
-\log\left(\cosh(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})\right)1_{\{|\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}|\geq r\}} \geq -\log\left(\cosh(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})\right),
$$

since $\cosh(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi}) \geq 1$ always holds. Using the inequality $\log\cosh(x) \leq |x|$ for x, the following bound holds

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\frac{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})}{1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right] \\
&\quad\geq \tfrac{1}{2}\log(2)\cdot\mathbb{E}_{\psi}\left[1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right]+\tfrac{1}{2}\mathbb{E}_{\psi}\left[\log\left|\sinh(\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{2})\right|1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right]-\tfrac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[\log(\cosh(\tfrac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi}))\right] \\
&\quad\geq \tfrac{1}{2}\log(2)\cdot\mathbb{E}_{\psi}\left[1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] \\
&\quad+\tfrac{1}{2}\mathbb{E}_{\psi}\left[\log\left|\sinh(\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{2})\right|1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right]-\tfrac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[|\tfrac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi}|\right] \\
&\quad\geq \tfrac{1}{2}\log(2)\cdot\mathbb{E}_{\psi}\left[1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] \\
&\quad+\tfrac{1}{2}\mathbb{E}_{\psi}\left[\log\left|\sinh(\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{2})\right|1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right]-\frac{3}{4}\left(\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}+\sqrt{\frac{2}{\pi}}\cdot\frac{\|\boldsymbol{\mu}\|}{\sigma}\right),
\end{aligned}
\qquad (3.59)
$$

where the last inequality uses (2.13) and $\rho \leq \frac{3}{\sigma^2}$. Using the inequality $|\sinh(x)| \geq \exp(\frac{|x|}{2})$ for all $|x| \geq 2\log(\sqrt{2}+1)$ and letting $r = 4\log(\sqrt{2}+1)$, we obtain

$$\begin{aligned}
\tfrac{1}{2}\log(2)\cdot \mathbb{E}_\psi\left[1_{\{|\psi|\geq \frac{4\log(\sqrt{2}+1)\}}{\sigma\|\tilde{\boldsymbol{\theta}}\|}}\right] + \tfrac{1}{2}\mathbb{E}_\psi\left[\log\left|\sinh(\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{2})\right| 1_{\{|\psi|\geq \frac{4\log(\sqrt{2}+1)}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] \\
&\geq \tfrac{1}{2}\log(2)\cdot \mathbb{E}_\psi\left[1_{\{|\psi|\geq \frac{4\log(\sqrt{2}+1)}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] + \tfrac{1}{2}\mathbb{E}_\psi\left[\left|\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{4}\right| 1_{\{|\psi|\geq \frac{4\log(\sqrt{2}+1)}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] \\
&\geq \tfrac{1}{2}\log(2)\cdot \mathbb{E}_\psi[1_{\{|\psi|\geq 1\}}] + \tfrac{1}{2}\mathbb{E}_\psi\left[\left|\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{4}\right| 1_{|\psi|\geq 1\}}\right] \qquad (3.60) \\
&\geq \left(\tfrac{1}{2}\log(2) + \frac{\sigma\|\tilde{\boldsymbol{\theta}}\|}{8}\right)\cdot \Phi^c(1).
\end{aligned}$$

Here (3.60) follows from the assumption that $\sigma\|\tilde{\boldsymbol{\theta}}\| \geq 436$. Combining (3.59), (3.60) and the bounds $\sigma \geq 0.33\|\boldsymbol{\mu}\|$ and $\sigma\|\tilde{\boldsymbol{\theta}}\| \geq 436$ the result follows.

Hinge loss. We begin by denoting $\xi_1 := \boldsymbol{\xi}^T\tilde{\boldsymbol{\theta}}$ and $\xi_2 := \boldsymbol{\xi}^T\boldsymbol{\mu}$. Notice that ξ_1 and ξ_2 are independent random variables. Recall that $\ell(t) := \ell(t,1) = \max(0, 1-t)$. We have that

$$\begin{aligned}
f(\boldsymbol{\theta}) - f(\rho\boldsymbol{\mu}) &= \mathbb{E}_{\boldsymbol{\xi}}\left[\ell(\boldsymbol{\xi}^T\boldsymbol{\theta}) - \ell(\rho\boldsymbol{\xi}^T\boldsymbol{\mu})\right] \\
&= \mathbb{E}_{\xi_1,\xi_2}\left[\ell(\xi_1 + \rho\xi_2) - \ell(\rho\xi_2)\right] \\
&= \mathbb{E}_{\xi_1,\xi_2}\left[\ell(-\xi_1 + \rho\xi_2) - \ell(\rho\xi_2)\right].
\end{aligned}$$

The second equality follows since $\xi_1 \sim -\xi_1$. We define the function

$$\kappa(\xi_1,\xi_2) := \ell(\xi_1 + \rho\xi_2) + \ell(-\xi_1 + \rho\xi_2) - 2\ell(\rho\xi_2).$$

We therefore obtain that

$$2\left(f(\boldsymbol{\theta}) - f(\rho\boldsymbol{\mu})\right) = \mathbb{E}_{\xi_1,\xi_2}\left[\kappa(\xi_1,\xi_2)\right].$$

Next we claim that

$$\kappa(\xi_1,\xi_2) = 0 \text{ whenever } |\xi_1| \leq |1-\rho\xi_2|. \qquad (3.61)$$

To see this, suppose that $|\xi_1| \leq |1-\rho\xi_2|$ holds. We consider two cases. First, assume that $0 \leq 1-\rho\xi_2$ which yields that $\rho\xi_2 - \xi_1 \leq 1$ and $\rho\xi_2 + \xi_1 \leq 1$. We therefore have $\kappa(\xi_1,\xi_2) = 1-\xi_1-\rho\xi_2+1+\xi_1-\rho\xi_2-2(1-\rho\xi_2) = 0$. Second, assume that $1-\rho\xi_2 \leq 0$. It thus holds that $1 \leq \rho\xi_2 - \xi_1$ and $1 \leq \rho\xi_2 + \xi_1$. Now it immediately follows that $\kappa(\xi_1,\xi_2) = 0$ and equation (3.61) is established. We claim the following

$$\kappa(\xi_1,\xi_2) = |\xi_1| - |1-\rho\xi_2| \text{ whenever } |\xi_1| \geq |1-\rho\xi_2|. \qquad (3.62)$$

To this end, we again consider two cases. First, assume that $\xi_1 \leq -|1-\rho\xi_2|$. This yields that $1 \leq -\xi_1 + \rho\xi_2$ and $\xi_1 + \rho\xi_2 \leq 1$, so it holds that $\kappa(\xi_1, \xi_2) = 1 - \xi_1 - \rho\xi_2 - 2\ell(\rho\xi_2)$. The claim (3.62) follows from the following simple identity

$$2\ell(t) = 1 - t + |1-t|, \quad \forall t \in \boldsymbol{R}. \tag{3.63}$$

Second, assume that $\xi_1 \geq |1-\rho\xi_2|$. It then holds that $\xi_1 + \rho\xi_2 \geq 1$ and $-\xi_1 + \rho\xi_2 \leq 1$ and therefore $\kappa(\xi_1, \xi_2) = 1 + \xi_1 - \rho\xi_2 - 2\ell(\rho\xi_2)$. The claim (3.62) follows from the identity (3.63). We therefore obtain

$$\begin{aligned}
\mathbb{E}_{\xi_1,\xi_2}[\kappa(\xi_1,\xi_2)] &= 2\mathbb{E}_{\xi_1,\xi_2}[(\ell(\xi_1+\rho\xi_2) + \ell(-\xi_1+\rho\xi_2) - 2\ell(\rho\xi_2))1_{\{\xi_1>0\}}] && (3.64)\\
&= 2\mathbb{E}_{\xi_1,\xi_2}[(\ell(\xi_1+\rho\xi_2) + \ell(-\xi_1+\rho\xi_2) - 2\ell(\rho\xi_2))1_{\{\xi_1\geq|1-\rho\xi_2|\}}] && (3.65)\\
&= \mathbb{E}_{\xi_1,\xi_2}[(\xi_1 - |1-\rho\xi_2|)1_{\{\xi_1\geq|1-\rho\xi_2|\}}]. && (3.66)
\end{aligned}$$

Here equation (3.64) holds because $\xi_1 \sim -\xi_1$ and $\kappa(\xi_1, \xi_2) = \kappa(-\xi_1, \xi_2)$. Equation (3.65) is true because of claim (3.61) and (3.66) follows from claim (3.62). From (3.66), we conclude that $\mathbb{E}_{\xi_1,\xi_2}[\kappa(\xi_1, \xi_2)] \geq 0$. We then observe the bound

$$\begin{aligned}
\mathbb{E}_{\xi_1,\xi_2}[(\xi_1 - |1-\rho\xi_2|)1_{\{\xi_1\geq|1-\rho\xi_2|\}}] &= \tfrac{1}{2}\mathbb{E}_{\xi_1,\xi_2}[\xi_1 - |1-\rho\xi_2| + |\xi_1 - |1-\rho\xi_2||]\\
&\geq -\tfrac{1}{2}\mathbb{E}_{\xi_2}[|1-\rho\xi_2|] + \tfrac{1}{2}\mathbb{E}_{\xi_1,\xi_2}[|\xi_1| - |1-\rho\xi_2|] && (3.67)\\
&= \tfrac{1}{2}\mathbb{E}_{\xi_1}[|\xi_1|] - \mathbb{E}_{\xi_2}[|1-\rho\xi_2|].
\end{aligned}$$

The second inequality follows from $\mathbb{E}_{\xi_1}[\xi_1] = 0$ and the triangle inequality $|x|-|y| \leq ||x|-y|$. On the other hand, it holds that

$$\mathbb{E}_{\xi_1}[|\xi_1|] = \sqrt{\frac{2}{\pi}} \cdot \sigma\|\tilde{\boldsymbol{\theta}}\|, \tag{3.68}$$

and

$$\mathbb{E}_{\boldsymbol{\xi}}[|1-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}|] \leq 1 + \rho\mathbb{E}_{\boldsymbol{\xi}}[|\boldsymbol{\mu}^T\boldsymbol{\xi}|] \leq 1 + \rho\|\boldsymbol{\mu}\|\left(\sqrt{\frac{2}{\pi}} \cdot \sigma + \|\boldsymbol{\mu}\|\right). \tag{3.69}$$

Combing equations (3.61), (3.62), (3.67), (3.68), and (3.69), we deduce

$$f(\boldsymbol{\theta}) - f(\rho\boldsymbol{\mu}) \geq \frac{1}{2}\left(\sqrt{\frac{1}{2\pi}} \cdot \sigma\|\tilde{\boldsymbol{\theta}}\| - 1 - \rho\|\boldsymbol{\mu}\|\left(\sqrt{\frac{2}{\pi}} \cdot \sigma + \|\boldsymbol{\mu}\|\right)\right). \tag{3.70}$$

Using the bounds $\sigma\|\tilde{\boldsymbol{\theta}}\| \geq 8 + 10\rho^*\sigma^2$, $\sigma \geq 0.62\|\boldsymbol{\mu}\|$ and $\rho \leq \frac{3}{2}\rho^*$, the result follows from (3.70). □

We next derive a lower bound (3.56), Part (b). But, first we need a basic lemma from convex analysis.

Lemma 12. *Suppose that $g : \boldsymbol{R}_{\geq 0} \to \boldsymbol{R}$ is a convex function with a minimizer at $\rho^* > 0$. Assume that g is twice differentiable on the interval $[\frac{3}{4}\rho^*, \frac{5}{4}\rho^*]$ and there exists a constant $B > 0$ such that $g''(\rho) \geq B$ for all $\rho \in [\frac{3}{4}\rho^*, \frac{5}{4}\rho^*]$. Then it holds that*

$$g(\rho) - g(\rho^*) \geq \frac{\rho^* B}{8}|\rho - \rho^*| \quad \text{for all } \rho \notin [\tfrac{1}{2}\rho^*, \tfrac{3}{2}\rho^*]. \tag{3.71}$$

Proof. The proof follows by considering the second order Taylor series expansion of the function g. □

Lemma 13. *(Lower bound for (b) in (3.56)) Fix $\boldsymbol{\theta} \in \boldsymbol{R}^d$ and orthogonally decompose $\boldsymbol{\theta} = \rho\boldsymbol{\mu} + \tilde{\boldsymbol{\theta}}$. Suppose that $|\rho - \rho^*| \geq \frac{1}{2}\rho^*$. Then provided that $\sigma \geq c\|\boldsymbol{\mu}\|$ where the constant c is defined in (3.35) and (3.36), there exists a positive constant A such that the following is true*

$$f(\rho\boldsymbol{\mu}) - f(\boldsymbol{\theta^*}) \geq A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}. \tag{3.72}$$

Proof. We consider the logistic and hinge loss separately.

Logistic loss. Define the function

$$g(\rho) := \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1 + \exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)\right], \quad \boldsymbol{\xi} \sim N(\boldsymbol{\mu}, \sigma^2 I_d).$$

By Lemma 6, we know that g is a convex function with a unique minimizer at $\rho^* := \frac{2}{\sigma^2}$. Observe that $f(\rho\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*) = g(\rho) - g(\rho^*)$; hence in order to prove (3.72), we instead aim to bound this difference in the function g. From (2.10), we have $\boldsymbol{\mu}^T\boldsymbol{\xi} \sim N(\|\boldsymbol{\mu}\|^2, \sigma^2\|\boldsymbol{\mu}\|^2)$. It thus holds

$$4g''(\rho) = \mathbb{E}\left(\frac{(\boldsymbol{\mu}^T\boldsymbol{\xi})^2}{\cosh(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})^2}\right) = \frac{1}{\sigma\|\boldsymbol{\mu}\|\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{z^2}{\cosh^2(\frac{\rho z}{2})}\exp\left(-\frac{(z - \|\boldsymbol{\mu}\|^2)^2}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right)dz.$$

Upper bounding $\cosh^2(\frac{\rho z}{2})$ by $\exp(|\rho z|)$, we next obtain

$$\begin{aligned}
4g''(\rho) &\geq \frac{1}{\sigma\|\boldsymbol{\mu}\|\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \exp\left(-|\rho z|\right) \exp\left(-\frac{(z-\|\boldsymbol{\mu}\|^2)^2}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right) dz \\
&= \frac{1}{\sigma\|\boldsymbol{\mu}\|\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \exp\left(-\frac{(z-\|\boldsymbol{\mu}\|^2)^2 + 2\sigma^2\|\boldsymbol{\mu}\|^2|\rho z|}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right) dz, \\
&= \frac{1}{\sigma\|\boldsymbol{\mu}\|\sqrt{2\pi}} \cdot \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} z^2 \exp\left(-\frac{z^2 - 2\|\boldsymbol{\mu}\|^2 z + 2\sigma^2\|\boldsymbol{\mu}\|^2|\rho z|}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right) dz \\
&= \frac{\sigma^2\|\boldsymbol{\mu}\|^2}{\sqrt{2\pi}} \cdot \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right) \int_{-\infty}^{+\infty} z^2 \exp\left(-\frac{z^2 - 2\frac{\|\boldsymbol{\mu}\|}{\sigma} z + 2|\rho z|\sigma\|\boldsymbol{\mu}\|}{2}\right) dz \\
&\geq \frac{\sigma^2\|\boldsymbol{\mu}\|^2}{\sqrt{2\pi}} \cdot \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right) \int_{0}^{+\infty} z^2 \exp\left(-\frac{z^2}{2}\right) \exp\left(z\left(\frac{\|\boldsymbol{\mu}\|}{\sigma} - \rho\sigma\|\boldsymbol{\mu}\|\right)\right) dz \\
&\geq \frac{\sigma^2\|\boldsymbol{\mu}\|^2}{\sqrt{2\pi}} \cdot \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2} - \frac{1}{2} - \left|\frac{\|\boldsymbol{\mu}\|}{\sigma} - \rho\sigma\|\boldsymbol{\mu}\|\right|\right) \int_{0}^{1} z^2 dz.
\end{aligned}$$

Here the second to last inequality follows from the change of variables $z \to z\sigma\|\boldsymbol{\mu}\|$. The last inequality follows from restricting the integral's domain to $[0,1]$ and also lower bounding $-\frac{z^2}{2}$ and $z\left(\frac{\|\boldsymbol{\mu}\|}{\sigma} - \rho\sigma\|\boldsymbol{\mu}\|\right)$ by $\frac{-1}{2}$ and $-\left|\frac{\|\boldsymbol{\mu}\|}{\sigma} - \rho\sigma\|\boldsymbol{\mu}\|\right|$ respectively. We see that

$$\exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2} - \frac{1}{2} - \left|\frac{\|\boldsymbol{\mu}\|}{\sigma} - \rho\sigma\|\boldsymbol{\mu}\|\right|\right) \geq \exp\left(-\frac{1}{2c^2} - \frac{1}{4c} - \frac{1}{2}\right),$$

for $\rho \in [\frac{3}{4}\rho^*, \frac{5}{4}\rho^*]$. By Lemma 12, the result follows with the constant A computed as follows

$$A = \frac{1}{12\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2c^2} - \frac{1}{4c} - \frac{1}{2}\right).$$

Hinge loss. We begin by defining the function $h(\rho) = f(\rho\boldsymbol{\mu})$. Therefore

$$f(\rho\boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\xi}}[\ell(\rho\boldsymbol{\xi}^T\boldsymbol{\mu})] = \mathbb{E}_{\boldsymbol{\xi}}[(1-\rho\boldsymbol{\xi}^T\boldsymbol{\mu})1_{\{\rho\boldsymbol{\xi}^T\boldsymbol{\mu}\leq 1\}}].$$

Hence, it holds that

$$h'(\rho) = \boldsymbol{\mu}^T\nabla f(\rho\boldsymbol{\mu}) = -\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\xi}^T\boldsymbol{\mu}1_{\{\rho\boldsymbol{\xi}^T\boldsymbol{\mu}\leq 1\}}].$$

From (2.10), we obtain that $\boldsymbol{\mu}^T\boldsymbol{\xi} \sim N(\|\boldsymbol{\mu}\|^2, \sigma^2\|\boldsymbol{\mu}\|^2)$. For $\rho > 0$, therefore, it holds that

$$h'(\rho) = \frac{-1}{\sigma\|\boldsymbol{\mu}\|\sqrt{2\pi}} \int_{-\infty}^{\frac{1}{\rho}} z \exp\left(-\frac{1}{2} \cdot \left(\frac{z}{\sigma\|\boldsymbol{\mu}\|} - \frac{\|\boldsymbol{\mu}\|}{\sigma}\right)^2\right) dz. \tag{3.73}$$

Applying chain rule thus yields

$$h''(\rho) = \frac{1}{\rho^3 \sigma \|\boldsymbol{\mu}\| \sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot \left(\frac{1}{\rho \sigma \|\boldsymbol{\mu}\|} - \frac{\|\boldsymbol{\mu}\|}{\sigma} \right)^2 \right) \quad \text{for all } \rho > 0.$$

Hence, for all $\rho \in [\frac{3}{4}\rho^*, \frac{5}{4}\rho^*]$ it holds that

$$h''(\rho) \geq \frac{64}{125 {\rho^*}^3 \sigma \|\boldsymbol{\mu}\| \sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot \Gamma^2 \right),$$

where $\Gamma := \max\left\{ \left| \frac{4}{3\rho^* \sigma \|\boldsymbol{\mu}\|} - \frac{\|\boldsymbol{\mu}\|}{\sigma} \right|, \left| \frac{4}{5\rho^* \sigma \|\boldsymbol{\mu}\|} - \frac{\|\boldsymbol{\mu}\|}{\sigma} \right| \right\}$. Therefore, by Lemma 12 and $|\rho - \rho^*| \geq \frac{1}{2}\rho^*$, it holds that

$$f(\rho\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*) \geq \frac{4}{125\sqrt{2\pi}} \cdot \frac{\sigma}{r\|\boldsymbol{\mu}\|} \cdot \exp\left(-\frac{1}{2} \cdot \Gamma^2 \right). \tag{3.74}$$

Here $r = \rho^* \sigma^2$. Note that $r > 0$ by Lemma 6. We aim to lower bound the right-hand side of (3.74). We denote by $w = \frac{\sigma}{r\|\boldsymbol{\mu}\|} - \frac{\|\boldsymbol{\mu}\|}{\sigma}$ the quantity defined in Lemma 6. In particular, by Lemma 6, the following holds

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{\sigma}{\|\boldsymbol{\mu}\|} = \Phi(w) \cdot \exp(\tfrac{1}{2}w^2). \tag{3.75}$$

We consider two cases. First suppose that $w \geq \frac{1}{(3\sqrt{2}-4)c}$. Along with the assumption $\frac{\sigma}{\|\boldsymbol{\mu}\|} \geq c$ this implies that $w \geq \frac{1}{3\sqrt{2}-4} \cdot \frac{\|\boldsymbol{\mu}\|}{\sigma}$. A simple computation shows that

$$w^2 \geq \frac{1}{2} \cdot \Gamma^2 \text{ for all } w \geq \frac{1}{3\sqrt{2}-4} \cdot \frac{\|\boldsymbol{\mu}\|}{\sigma}$$

On the other hand, by (3.75) for $w \geq 0$, we obtain that $\frac{2}{\pi} \cdot \frac{\sigma^2}{\|\boldsymbol{\mu}\|^2} \geq \exp(w^2)$. Plugging in the bounds

$$w^2 \geq \frac{1}{2} \cdot \Gamma^2, \exp(-w^2) \geq \frac{\pi}{2} \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}, \text{ and } \frac{\sigma}{r\|\boldsymbol{\mu}\|} \geq w \geq \frac{1}{(3\sqrt{2}-4)c}$$

into the right-hand-side of (3.74), we obtain that

$$f(\rho\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*) \geq \frac{\sqrt{2\pi}}{125(3\sqrt{2}-4)c} \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}.$$

Next, suppose that $w < \frac{1}{(3\sqrt{2}-4)c}$. In this case, the two factors $\frac{\sigma}{r\|\boldsymbol{\mu}\|}$ and $\exp\left(-\frac{1}{2}\cdot\Gamma^2\right)$ in (3.74) are lower bounded separately. Note that it always holds that $w \geq -\frac{\|\boldsymbol{\mu}\|}{\sigma}$ as $r > 0$. Therefore, it is easy to see that the latter factor is lower bounded by

$$\exp\left(-\frac{1}{2}\left(\frac{4}{3(3\sqrt{2}-4)c}+\frac{1}{3c}\right)^2\right).$$

Hence, it remains to bound the factor $\frac{\sigma}{r\|\boldsymbol{\mu}\|}$ in (3.74). To this end, we show that $w \geq -\frac{\|\boldsymbol{\mu}\|}{2\sigma}$ for all $\frac{\sigma}{\|\boldsymbol{\mu}\|} \geq c$. Note that a chain of change of variables gives

$$\Phi(w)\cdot\exp\left(\frac{w^2}{2}\right) = \frac{1}{\sqrt{2\pi}}\cdot\int_0^{+\infty}\exp(-\frac{1}{2}t^2)\cdot\exp(wt)\,dt.$$

The right-hand side of (3.75) is an increasing function with respect to w. Therefore it suffices to show that the following holds

$$\frac{1}{\sqrt{2\pi}}\cdot\frac{\sigma}{\|\boldsymbol{\mu}\|} \geq \Phi\left(-\frac{\|\boldsymbol{\mu}\|}{2\sigma}\right)\cdot\exp\left(\frac{\|\boldsymbol{\mu}\|^2}{8\sigma^2}\right) \quad \text{whenever} \quad \frac{\sigma}{\|\boldsymbol{\mu}\|} \geq c. \tag{3.76}$$

However, it can be verified by a plot that

$$\tfrac{1}{\sqrt{2\pi}} \geq t\cdot\Phi\left(-\tfrac{t}{2}\right)\cdot\exp\left(\frac{t^2}{8}\right) \text{ holds for all } t \in (0,\frac{1}{c}).$$

Therefore, we have shown that $w \geq -\frac{\|\boldsymbol{\mu}\|}{2\sigma}$ which implies that $\frac{\sigma}{r\|\boldsymbol{\mu}\|} \geq \frac{\|\boldsymbol{\mu}\|}{2\sigma}$. Finally we lower bound the quantity $\frac{\sigma}{r\|\boldsymbol{\mu}\|}$ by $c\cdot\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}$. We have concluded (3.72) in case of hinge loss function where the constant A can be computed as follows

$$A = \min\left\{\frac{c}{2}\cdot\exp\left(-\frac{1}{2}\left(\frac{4}{3(3\sqrt{2}-4)c}+\frac{1}{3c}\right)^2\right), \frac{\sqrt{2\pi}}{125(3\sqrt{2}-4)c}\right\}.$$

□

We now have the ingredients to prove Theorem 7.

Proof of Theorem 7. Consider the set C and function V defined in (3.51):

$$C := \left\{\boldsymbol{\theta} : |\rho-\rho^*| < \tfrac{1}{2}\rho^* \text{ and } \sigma\|\tilde{\boldsymbol{\theta}}\| \leq c'\right\} \quad \text{and} \quad V(\boldsymbol{\theta}) = \frac{1}{2\alpha}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2. \tag{3.77}$$

We let c' to be defined as in Lemma 11. This means that c' equals to 436 and $8 + 10\rho^*\sigma^2$ in case of logistic and hinge loss respectively. We next show that there exists a positive constant δ such that the following is true

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} \geq 1\right) \geq \delta \quad \text{for all} \quad \boldsymbol{\theta} \in C. \tag{3.78}$$

Let $\boldsymbol{\theta} \in C$ and orthogonally decompose it into $\boldsymbol{\theta} = \rho\boldsymbol{\mu} + \tilde{\boldsymbol{\theta}}$. We have that $\boldsymbol{\xi}^T\boldsymbol{\theta} = \rho\boldsymbol{\xi}^T\boldsymbol{\mu} + \boldsymbol{\xi}^T\tilde{\boldsymbol{\theta}}$. Note that $\rho > 0$ as $\boldsymbol{\theta} \in C$. By (2.11), we see that $\boldsymbol{\xi}^T\boldsymbol{\theta}$ and $\boldsymbol{\xi}^T\tilde{\boldsymbol{\theta}}$ are independent normal random variables. It thus holds that

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} \geq 1\right) \geq \mathbb{P}_{\boldsymbol{\xi}}\left(\rho\boldsymbol{\xi}^T\boldsymbol{\mu} \geq 1\right) \cdot \mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\tilde{\boldsymbol{\theta}} \geq 0\right) = \frac{1}{2} \cdot \mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\boldsymbol{\mu} \geq \frac{1}{\rho}\right). \tag{3.79}$$

Rewrite the inequality $\boldsymbol{\xi}^T\boldsymbol{\mu} \geq \frac{1}{\rho}$ by $z := \frac{\boldsymbol{\xi}^T\boldsymbol{\mu} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|} \geq \frac{\frac{1}{\rho} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}$. Noting that $z \sim N(0,1)$ and using the inequality $\frac{2}{\rho^*} \geq \frac{1}{\rho}$, we obtain that

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} \geq 1\right) \geq \delta := \frac{1}{2} \cdot \Phi^c\left(\frac{\frac{2}{\rho^*} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}\right). \tag{3.80}$$

We next show that the pair (C, V) satisfies the drift equation (3.22). Let us rewrite (3.56):

$$f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) = \underbrace{f(\boldsymbol{\theta}_{k-1}) - f(\rho_{k-1}\boldsymbol{\mu})}_{(a)} + \underbrace{f(\rho_{k-1}\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*)}_{(b)}. \tag{3.81}$$

By Lemmas 11 and 13, both terms in (a) and (b) in (3.81) are non-negative . Assume that $\boldsymbol{\theta}_{k-1} \notin C$. Therefore, either $\sigma\|\tilde{\boldsymbol{\theta}}_{k-1}\| \geq c'$ or $|\rho_{k-1} - \rho^*| \geq \frac{1}{2}\rho^*$; this implies that the quantity (a) is at least 1 or the quantity (b) is at least $A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$ respectively. The constant A in Lemma 13 satisfies $1 \geq A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$ for all $\frac{\sigma}{\|\boldsymbol{\mu}\|} \geq c$. Hence it holds that

$$A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2} \leq f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) \quad \text{for all } \boldsymbol{\theta}_{k-1} \notin C. \tag{3.82}$$

We use (3.53) next to establish the drift equation (3.22). Recall that the following holds

$$f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) \leq \frac{1}{2\alpha}\left(\|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\|^2 - \mathbb{E}\left[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2 | \mathcal{F}_{k-1}\right]\right) + \frac{\alpha}{2}\left(\|\boldsymbol{\mu}\|^2 + d\sigma^2\right). \tag{3.83}$$

Combining the last two displayed inequalities and using the definition of function V, we obtain that

$$\left(\mathbb{E}\left[V(\boldsymbol{\theta}_k)|\mathcal{F}_{k-1}\right] - V(\boldsymbol{\theta}_{k-1})\right) \cdot 1_{\{\boldsymbol{\theta}_{k-1} \notin C\}} \leq \left(\frac{\alpha}{2}(\|\boldsymbol{\mu}\|^2 + d\sigma^2) - A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}\right) \cdot 1_{\{\boldsymbol{\theta}_{k-1} \notin C\}}. \tag{3.84}$$

Therefore, by choosing $\alpha < A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2(\|\boldsymbol{\mu}\|^2+d\sigma^2)}$, we obtain the drift equation (3.22) holds with $b := \frac{A}{2} \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$. Next, we obtain bounds on $\mathbb{E}[\tau_m]$ for $m \geq 1$. By Lemma 1 and a simple induction, we obtain that

$$\mathbb{E}[\tau_m] \leq \tfrac{1}{b}V(0) + \tfrac{1}{b}(m-1)\sup_{\boldsymbol{\theta}\in C} V(\boldsymbol{\theta}). \tag{3.85}$$

Compactness of set C yields that, $\sup_{\boldsymbol{\theta}\in C} V(\boldsymbol{\theta}) < +\infty$. Therefore, for some constant γ, the following is true

$$\mathbb{E}[\tau_m] \leq \gamma \cdot m. \tag{3.86}$$

Combining (3.86), (3.80) and Lemma 8, the proof immediately follows. □

3.3.3 Angle bound, proof of Theorem 8

Proof of Theorem 8. Recall the SGD algorithm for logistic regression uses the update

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \frac{\alpha\boldsymbol{\xi}_k}{1+\exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})}$$

and for hinge regression

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \alpha 1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}\boldsymbol{\xi}_{k-1}$$

where $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots \overset{i.i.d}{\sim} N(\boldsymbol{\mu}, \sigma^2 I_d)$. It clearly holds in both cases that

$$\big||\boldsymbol{v}^T\boldsymbol{\theta}_k| - |\boldsymbol{v}^T\boldsymbol{\theta}_{k-1}|\big| \leq \alpha|\boldsymbol{v}^T\boldsymbol{\xi}_{k-1}|. \tag{3.87}$$

We define a new random variable $X_k := |\boldsymbol{v}^T\boldsymbol{\theta}_k| - k\sigma\alpha\sqrt{\frac{2}{\pi}}$. Observe that $\mathbb{E}\left[|X_0|\right] = 0$ and for all $k \geq 1$, it holds that

$$\mathbb{E}\left[|X_k|\right] \leq \alpha\sum_{i=1}^{k}\mathbb{E}\left[\left|\boldsymbol{v}^T\boldsymbol{\xi}_k\right|\right] + k\sigma\alpha\sqrt{\frac{2}{\pi}} < \infty,$$

i.e., $X_k \in \mathcal{L}^1$ for all $k \geq 1$. Next, we have for any $k \geq 1$

$$\mathbb{E}\left[|X_k - X_{k-1}| \mid \mathcal{F}_{k-1}\right] \leq \mathbb{E}\left[\big||\boldsymbol{v}^T\boldsymbol{\theta}_k| - |\boldsymbol{v}^T\boldsymbol{\theta}_{k-1}|\big| \mid \mathcal{F}_{k-1}\right] + \sigma\alpha\sqrt{\frac{2}{\pi}} \leq 2\sigma\alpha\sqrt{\frac{2}{\pi}}.$$

Here we used that $\boldsymbol{v}^T\boldsymbol{\xi}_k \sim N(0,\sigma^2)$ along with (2.13). We also see that

$$\mathbb{E}\left[|\boldsymbol{v}^T\boldsymbol{\theta}_k|\,|\,\mathcal{F}_{k-1}\right] \leq |\boldsymbol{v}^T\boldsymbol{\theta}_{k-1}| + \sigma\alpha\sqrt{\frac{2}{\pi}} \quad \Rightarrow \quad \mathbb{E}\left[X_k\,|\,\mathcal{F}_{k-1}\right] \leq X_{k-1}.$$

Therefore, we have shown that $X_0, X_1, \cdots$ is a super-martingale. By Theorem 5, we have $\mathbb{E}\left[X_T\right] \leq 0$. The result follows.

□

3.4 Numerical experiments

We investigate the performance of our termination test on two popular data sets, MNIST [51] and CIFAR-10 [49], as well as synthetic data generated from Gaussians and heavy-tailed student t-distributions. All tests were performed using our zero overhead stopping criteria outlined in (3.15); experiments using our test which required an extra sample (3.14) are not presented since the behaviors of the two criteria were indistinguishable on all data sets.

Comparison with a popular stopping criterion. We include as a baseline a popular termination test, the small validation set (SVS) [72]. The SVS termination test is as follows. One fixes a validation set of p instances $(\boldsymbol{\zeta}_1^{\mathrm{V}}, y_1^{\mathrm{V}})$, $\ldots$, $(\boldsymbol{\zeta}_p^{\mathrm{V}}, y_p^{\mathrm{V}})$ drawn from the same distribution as the training data. Then for $m = 1, 2, \ldots$, one checks the fraction correct of the current classifier $\boldsymbol{\theta}_{ml}$, where ml is the iteration index, on the p instances. In other words, the SVS test is run once every l iterations. If the fraction correct fails to increase compared to the last run of the SVS, then the SGD iterations are terminated.

Note the computational overhead of running the small validation set is about p times the cost of one SGD iteration. Therefore, in order to make the overhead only a constant factor, we choose $l = 2p$, meaning an approximately 50% overhead for SVS. In contrast, the overhead for (3.15) is 0. The value of p is a tuning parameter for SVS; we exhibit results for three different p values (see Figs. 3.3, 3.5, 3.7, 3.9).

Measuring the accuracy. In all the experiments, we measure the performance of a method with a score, generally known as "accuracy," that is the fraction correct on a large validation set drawn from the same distribution as the training data. Thus, 1.0 is perfect accuracy, while 0.5 means that $\boldsymbol{\theta}_k$ is no better at classifying than random guessing. It is

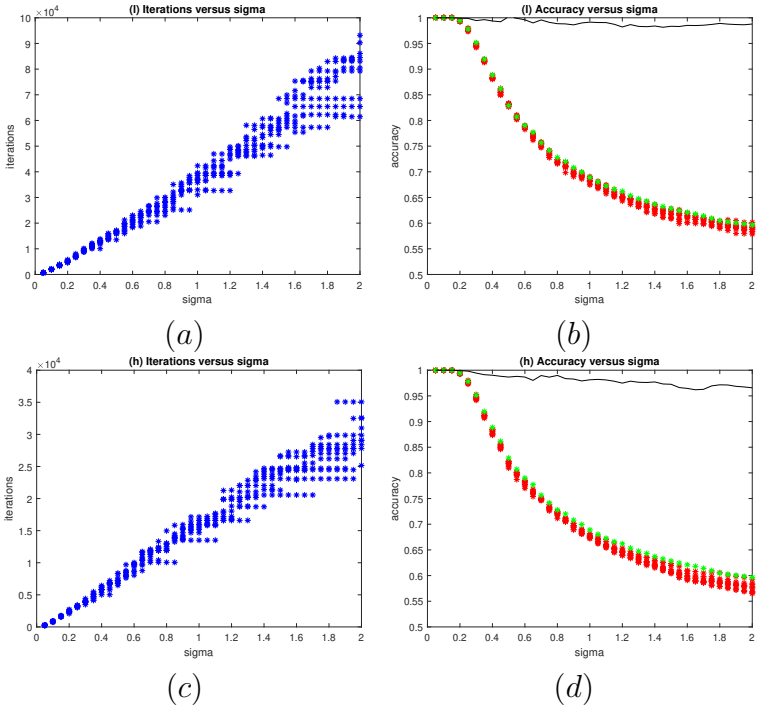

Figure 3.2: Performance of stopping criterion (3.15) on a mixture of Gaussians as σ is varied. Plots $(a), (b)$ are logistic and $(c), (d)$ are hinge. All plots show tests for values of σ equally spaced from 0.05 to 2.0. For each value of σ, ten trials were run. Plots $(a), (c)$ show the relationship between σ and k, the iteration number when (3.15) first holds. Plots $(b), (d)$ show the accuracy as red asterisks. The green asterisks show the accuracy of the optimal classifier. The black curve on the right is the ratio of the average accuracy (over 10 trials) of the classifier when (3.15) holds to the accuracy of the optimal classifier.

important to note that even on data for which the means $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ are known a priori (*e.g.*, synthetic data), the score of the optimal $\boldsymbol{\theta}^*$ will not be 1.0 because the large validation set itself is noisy.

We center the data so that the linear classifier is homogeneous. In a preliminary phase, 100 samples are drawn from the training set. From this, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are estimated, and then the average of these estimates is used to offset training instances during SGD.

Parameter settings. After centering, the vectors $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ scale inversely, so the step-size parameter α should scale as $1/\sigma^2$. Therefore, we take the step-size to be $\tilde{\alpha}/\tilde{\sigma}^2$. Here, $\tilde{\sigma}^2$ is the average of $\left\|\boldsymbol{\zeta}_j - \tilde{\boldsymbol{\mu}}_{y_j}\right\|^2$, and $\tilde{\boldsymbol{\mu}}_i$ ($i = 0$ or $i = 1$) is the estimate of $\boldsymbol{\mu}_i$, averaged over the two classes. We compute the quantities $\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\mu}}_i$ using the 100 samples described in the preceding paragraph. Note that for the Gaussian mixture model, the expected value of $\tilde{\sigma}^2$ is $\sigma^2 d$. For the synthetic data, the means and variances are known exactly a priori, so the estimation procedures described in the previous two paragraphs are unnecessary. However, we used them anyway in order to be consistent with the tests on the realistic data.

The parameter $\tilde{\alpha}$ described in the last paragraph is a scale-free tuning parameter. It is known (see, e.g., [61]) that a smaller $\tilde{\alpha}$ corresponds to more iterations but greater ultimate accuracy under a reasonable model of the data. Our termination test is obviously sensitive to the choice of $\tilde{\alpha}$: the condition $\boldsymbol{\xi}_{k+1}^T \boldsymbol{\theta}_k \geq 1$ cannot hold unless $\|\boldsymbol{\theta}_k\| \geq 1/\|\boldsymbol{\xi}_{k+1}\|$, but $\mathbb{E}[\|\boldsymbol{\theta}_k\|] \leq O(\alpha k)$. See also Theorems 6 and 7. On the other hand, SVS is only mildly sensitive to $\tilde{\alpha}$, according to our testing. Indeed, there is an upper bound of pl on the total number of iterations possible before termination using the SVS condition, independent of $\tilde{\alpha}$ and of all other aspects of the problem. The dependence of the termination test on $\tilde{\alpha}$ is evidently desirable because the user is presumably seeking greater accuracy when a smaller value of $\tilde{\alpha}$ is selected.

3.4.1 Experiments with synthetic data

Normal distribution. We generated test and training data using a mixture of Gaussians given by $N(\mathbf{0}, \sigma^2 I)$ for the 0-class and $N(\boldsymbol{e}_1, \sigma^2 I)$ for the 1-class, where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T \in \boldsymbol{R}^d$.

In Fig. 3.2, we present the running time and accuracy (fraction correct) of our termination test for a fixed dimension $d = 500$ and σ ranging from 0.05 to 2. We record 10 runs for each value of σ. The performance of the classifier when our termination test (3.15) holds almost matches the optimal classifier; in particular, the averaged accuracy of our

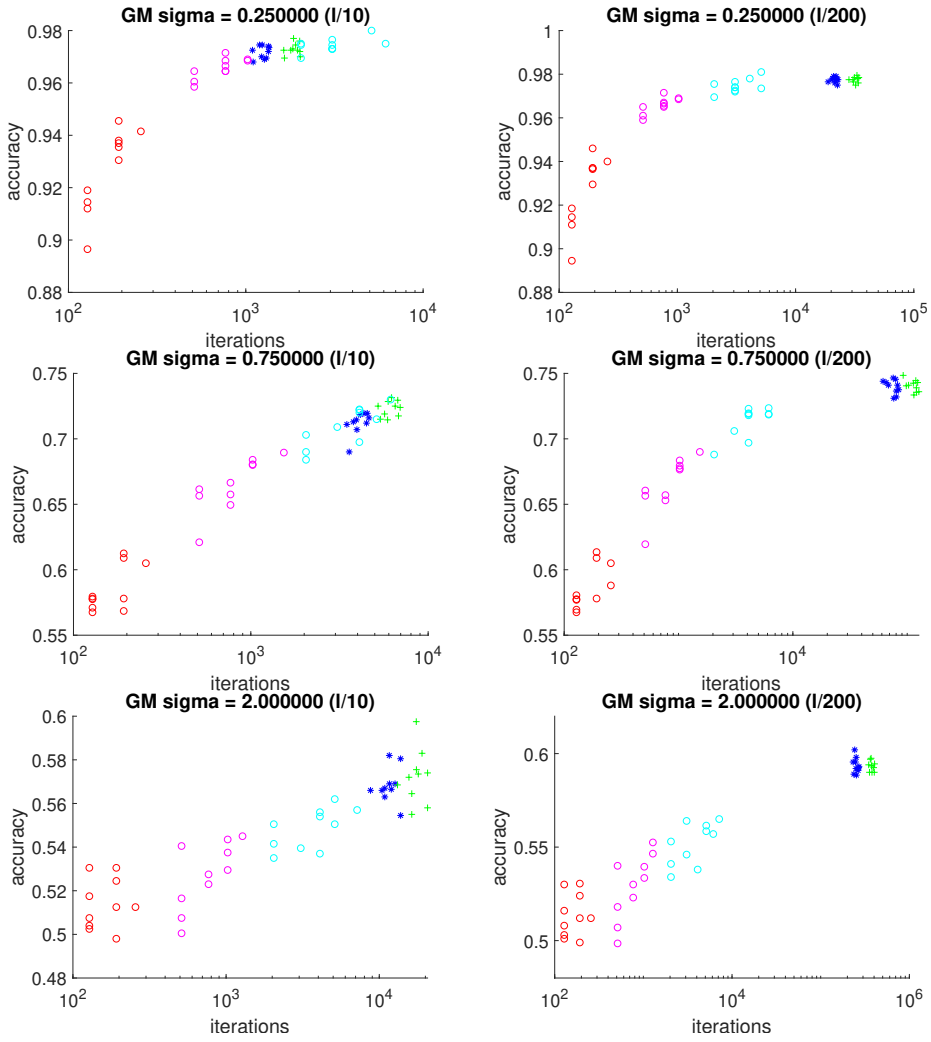

Figure 3.3: Each plot shows 10 random runs of SGD applied to normally distributed data with indicated values of σ and for a fixed dimension $d = 500$. For each of the ten runs, five termination tests corresponding to five colors were applied. SVS was tried with $p = 32, 128, 512$, depicted as red, magenta and cyan circles respectively. Test (3.15) is indicated with a blue asterisk. A green '+' corresponds to termination after $1.5k$ iterations, where k is the iteration index that (3.15) first holds. The notation $(l/200)$ means logistic loss with $\tilde{\alpha} = 1/200$; simillarly $(h/10)$ means hinge loss with $\tilde{\alpha} = 1/10$, and so on.

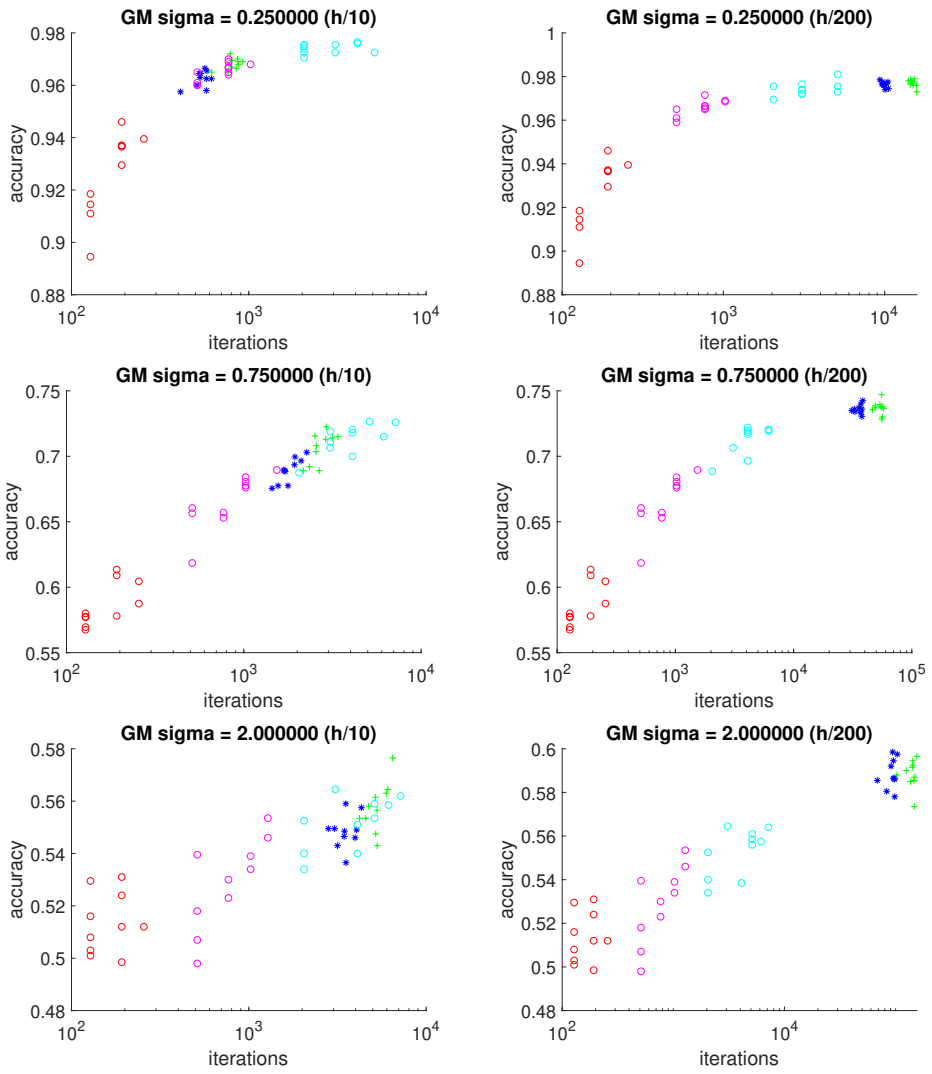

Figure 3.4: Refer to the caption of Fig. 3.3 for the key to the plots.

classifier/accuracy of the optimal classifier over the 10 runs, black curve in Fig. 3.4, never dips below 0.95.

In Fig. 3.3, we compare performance of (3.15) against SVS termination. One axis shows accuracy while the other shows iteration count. We continued to run SGD for an additional $1.5k$ iterations where k is the first iteration at which (3.15) holds (green '+') to test whether accuracy improves after termination. The tests (for several values of σ, both hinge and logistic, and two values of $\tilde{\alpha}$) in Fig. 3.3 indicate that (3.15) is more accurate than SVS, more predictable (i.e., there is less spread in the scatter plot), and that running until $1.5k$ iterations does not significantly improve the solution. As expected, for a large $\tilde{\alpha}$, (3.15) requires fewer iterations than SVS with $p = 512$, while the opposite relationship holds for a small $\tilde{\alpha}$.

Heavy-tailed distribution. We consider the student t-distribution with two degrees of freedom. This distribution is heavy-tailed since some of its higher moments are infinite.

The two classes were generated as follows. For $\boldsymbol{\zeta}$ in the 0-class, each of the d entries of $\boldsymbol{\zeta}$ is chosen as $\beta\eta$, where β is varied in the experiments and η is drawn from the student t-distribution with two degrees of freedom. For the 1-class, $\boldsymbol{\zeta}$ is chosen in the same way except that the first entry is incremented by 1. Fig. 3.5 shows our performance against SVS. The results in this table show similar trends as in the normally distributed case. One difference is that the accuracy achieved by our termination test (3.15) is more spread out presumably because of the heavy-tailed nature of the data set.

3.4.2 Experiments with real data

MNIST handwritten digits. We compared our termination test on the MNIST handwritten digit set [51] ($d = 784$, no preprocessing of the data other than centering between the two means). Two trials are shown: distinguishing 1 from 8 (easy case) and distinguishing 7 from 9 (more difficult case). The test runs are obtained by running through the training data in different randomized orders. The plots in Fig. 3.7 show similar trends as before. As expected, the accuracy is overall higher for $\tilde{\alpha} = 1/200$ than for $\tilde{\alpha} = 1/10$.

CIFAR-10 image set. We compared our termination test on the CIFAR-10 [49] ($d = 3072$, no preprocessing of the data other than centering between the two means as described earlier). Two trials are shown: distinguishing deer from airplanes and frogs from trucks. As in MNIST, test runs are obtained by running through the training data in different randomized orders.

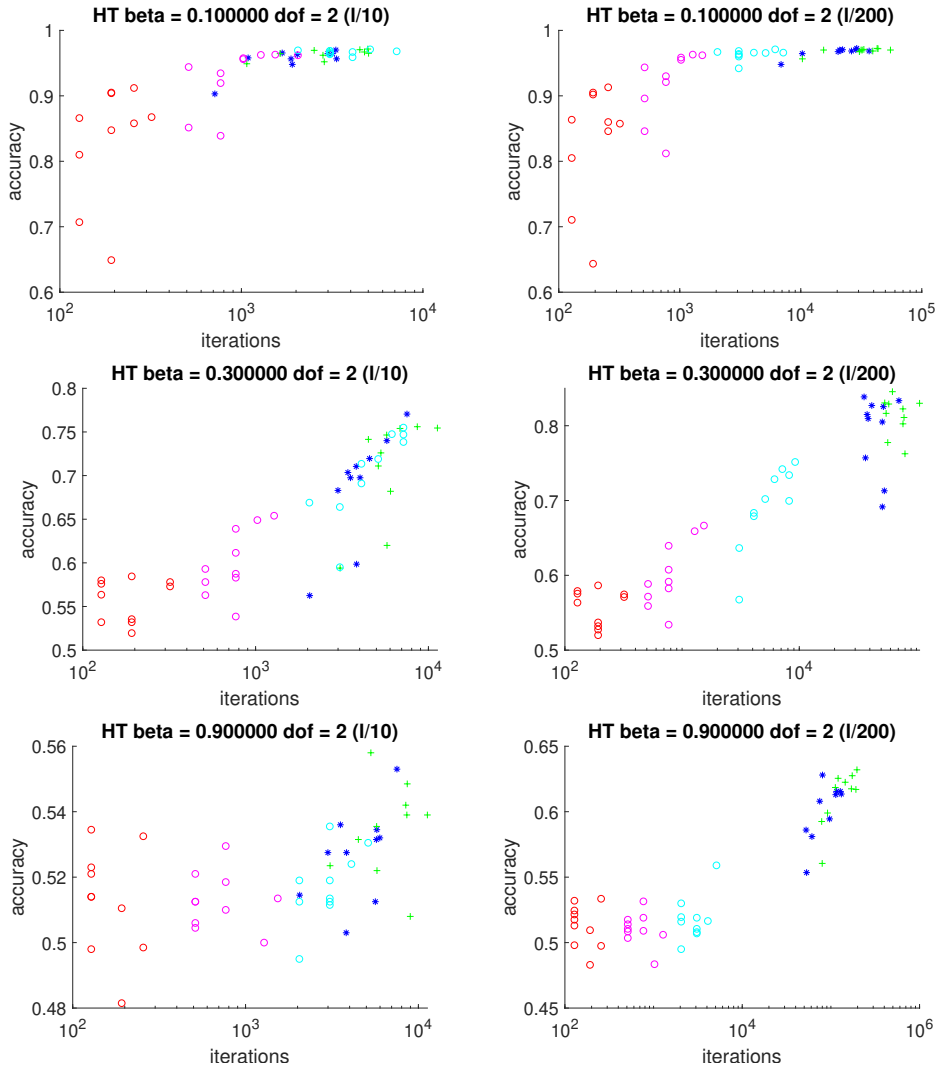

Figure 3.5: Tests on the student-t distribution (heavy tailed) with two degrees of freedom and the indicated value of parameter β. See the caption of Fig. 3.3 for explanation of the plots.

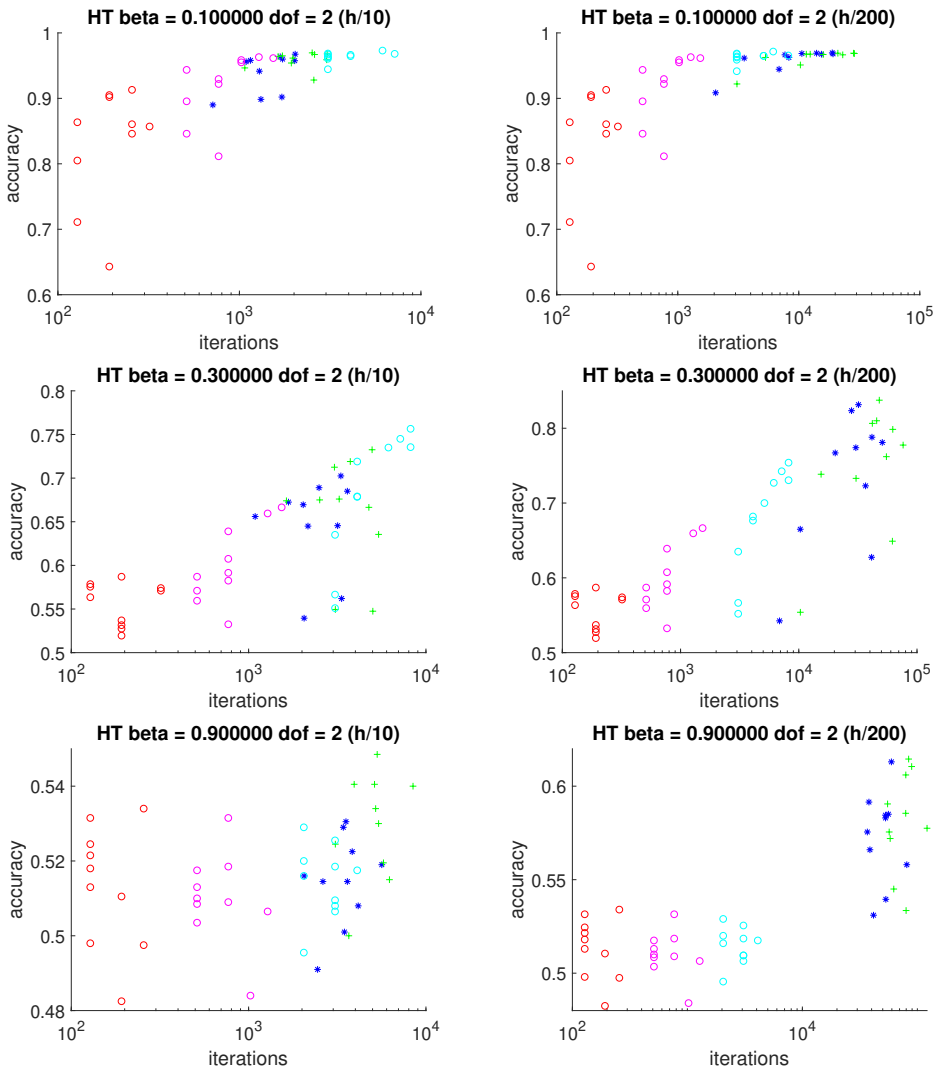

Figure 3.6: Refer to the caption of Fig. 3.5 for the key to the plots

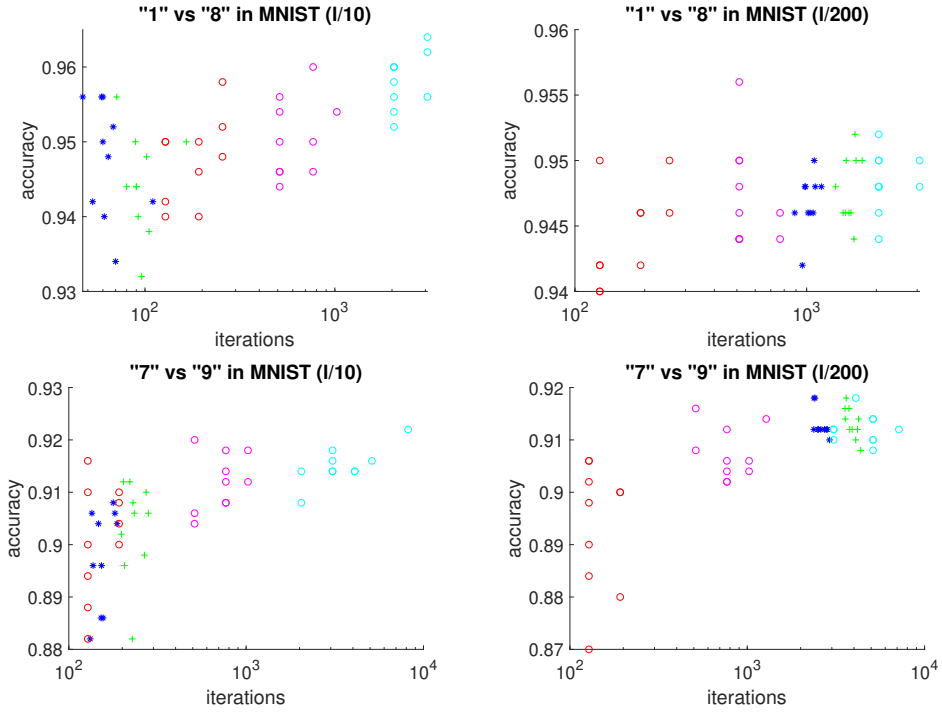

Figure 3.7: Tests on the MNIST handwritten digit data set for discerning "1" from "8" and "7" from "9" for both hinge and logistic, and for both $\tilde{\alpha} = 1/10$ and $\tilde{\alpha} = 1/200$. Refer to the caption of Fig. 3.3 for the key to the plots.

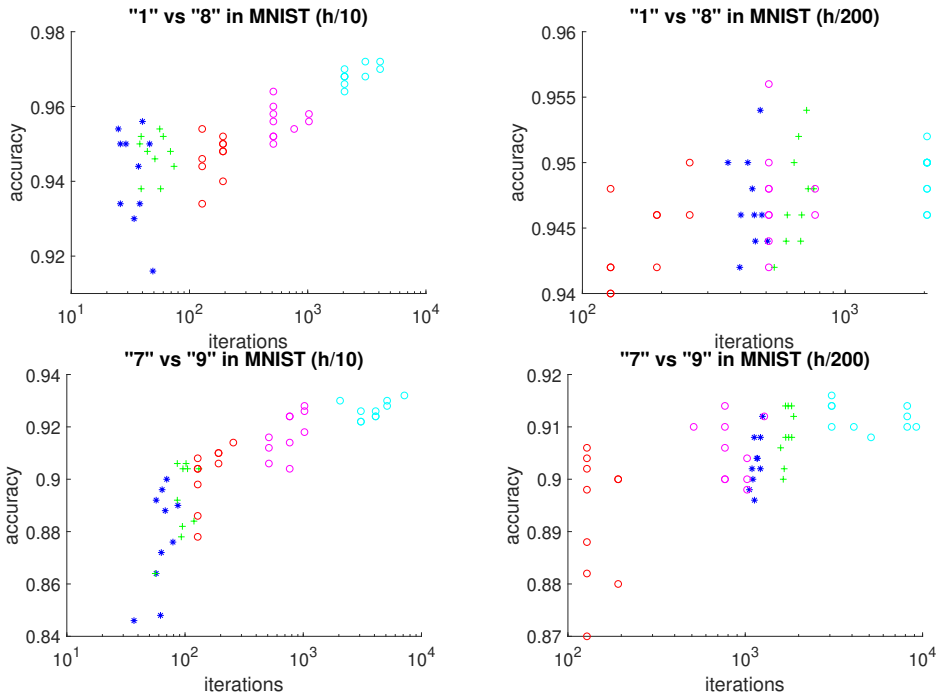

Figure 3.8: Refer to the caption of Fig. 3.7 for the key to the plots

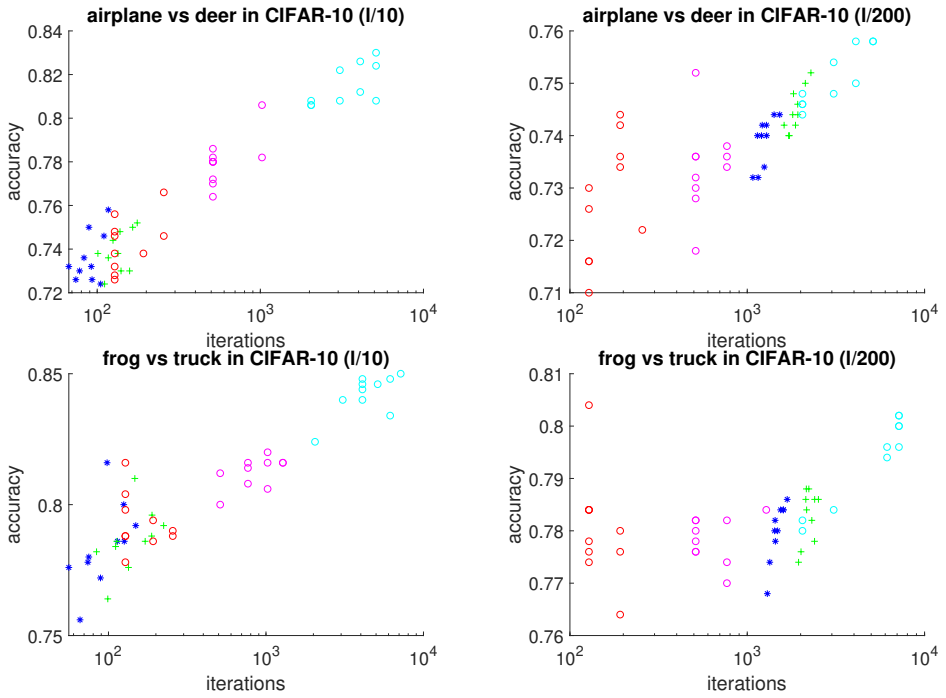

Figure 3.9: Tests on the CIFAR-10 image set for two tasks, for logistic and hinge losses, and for $\tilde{\alpha} = 1/10$ and $\tilde{\alpha} = 1/200$. Refer to the caption of Fig. 3.3 for the key to the plots. The plot in the first row, right, does not include cyan circles because the training data was exhausted before the SVS test could activate for $p = 512$.

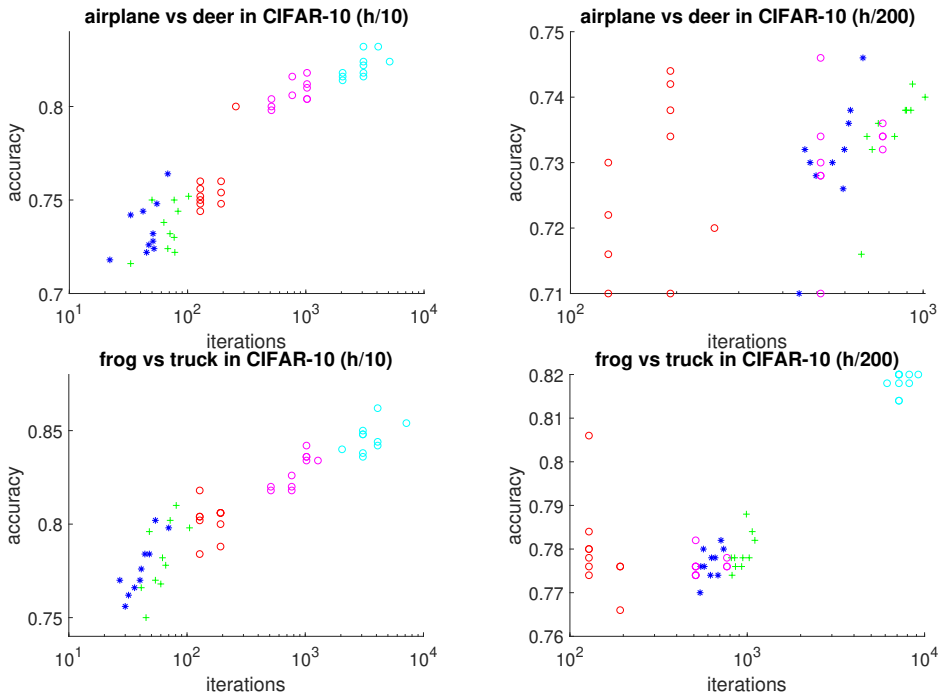

Figure 3.10: Refer to the caption of Fig. 3.9 for the key to the plots

Chapter 4

SGD with Early Stopping for Least Squares Deconvolution

Deconvolution has a wide range of application including image deblurring [34], electrical impedance tomography [42] and optical microscopy [46]. The main aim of this chapter is to explore the implicit regularization effect of the SGD algorithm with early stopping for the least squares deconvolution problem for the task of image deblurring.

The key results of this chapter are as follows.

- Motivated by experimental observations, the SGD least squares deconvolution algorithm exhibits the following phenomenon: SGD converges to a vicinity of the ground solution (sharp image) after only a few batches of iterations. In other words, SGD with early stopping has an implicit regularization effect.
- Theoretical justification is provided to ensure that SGD, with high probability, shall follow the gradient flow trajectory (Theorem 9). Our approach is novel in the sense that we establish our results by way of concentration of measures.
- Motivated by numerical evidence, we propose a new stopping time for SGD which can be easily implemented for both GD and SGD algorithms. Based on the fact that both GD and SGD behave similarly and that the GD algorithm is more amenable to analysis, we analyze our stopping time for the GD algorithm instead of SGD (Theorem 10).

- We conclude with a new concentration inequality for products of random contractions (Theorem 11) which can be helpful in analysing other stochastic algorithms such as k-streaming PCA [44].

The outline of this chapter is as follows: Section 4.1 is a brief description of image deblurring problems. Section 4.2 introduces and describes the least squares deconvolution problem. In Section 4.3, we first recall that GD with early stopping does exhibit implicit regularization. Next, we provide a theoretical result (Theorem 9) asserting that the SGD least square deconvolution algorithm behaves similarly to GD. In Section 4.4, we present our experimental observations. It is emphasized that a full experimental study of the SGD algorithm for the deconvolution least squares problem is outside of the scope of this thesis; our numerical experiment in this chapter is only for the purpose of motivating our analysis. In Section 4.5, we propose our new stopping time and establish a theoretical upper bound for the error term at termination. Finally, in Section 4.6, we establish a new concentration bound for products of random contractions.

4.1 Image deblurring

Image deblurring is an important class of inverse problems, which are mathematical problems that arise when our aim is to reconstruct from a set of observations the causal factors that created them. See Figure 4.1. These problems have wide application in medical imaging, computer vision and machine learning, etc.

It is well known what a blurred image looks like and why they are visually unappealing. Blurring is the operation (system) of producing the blurred image (output) from the sharp image (input), occurs for diverse reasons, including defocusing camera's lens, motion blur caused by the object not being still when camera's shutter was open, or due to variations in the air that impact the light coming into the camera. In image deblurring, we are given a blurred image and our task is to recover the sharp image. In order to reconstruct the sharp image, we take a model-based approach where the blurring process is described based on a concise mathematical model. The mathematical techniques used for image reconstruction are called deconvolution methods.

Before discussing the way in which the blurring process is modeled, we first require to represent an image using some mathematical object. A digital image (color or grayscale) is composed of pixels (picture elements), each with its own corresponding light intensity. In grayscale images, each pixel's intensity is quantified by some integer value inside the interval $[0, 255]$ where 0 and 255 indicates black and white respectively. Based on this

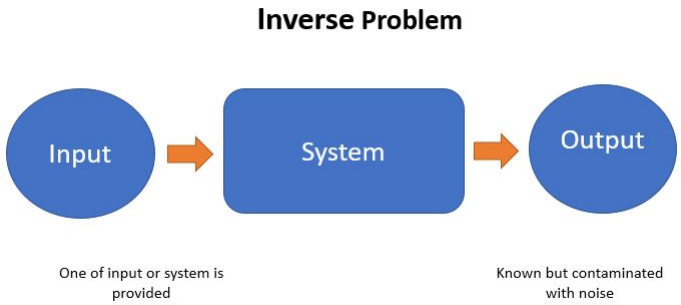

Figure 4.1: The inverse problem is to reconstruct the system or the input while the other two quantities are provided. Almost always, the output is revealed to us imprecisely meaning that it has been contaminated with some noise.

representation, we shall present a grayscale image as a rectangular matrix whose elements are its pixels' intensities. After that, by stacking the columns of this matrix[1], we obtain a vector. All in all, we can represent the sharp image and blurred image by real-valued vectors $\boldsymbol{x}^*$ and $\boldsymbol{b}$ respectively.

We now state the assumption most commonly made about the blurring process. This assumption is widely believed to be realistic in physical sciences and enables us to use a large variety of mathematical tool-kits to perform the deblurring process.

Assumption 2. The blurring process is assumed to be linear, *i.e.,* there exists a blurring matrix A such that $\boldsymbol{b} = A\boldsymbol{x}^* + \boldsymbol{\xi}$. Here $\boldsymbol{\xi}$ represents the additive noise in the observed blurred image.

As in Assumption 2, we denote the additive noise in the blurred image by a vector $\boldsymbol{\xi}$. In this chapter, we use the Gaussian white noise model which is the most used model in image restoration literature.

Assumption 3. We assume that the additive noise $\boldsymbol{\xi}$ is drawn from a isotropic Gaussian distribution with mean zero. In other words, we let $\boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2 I_m)$ where m is the dimension of $\boldsymbol{b}$ and the standard deviation σ is proportional to the amplitude of the noise.

To complete our model, we still need to explain how we obtain the blurring matrix A. To this end, we perform the following thought experiment: consider an image where all the

[1]The mathematical symbol for this operation is denoted by vec.

pixels are pitch black (intensity of 0) except only a single bright pixel (intensity of 255). Taking a picture from this image will reveal to us how the blurring operator causes the bright pixel to *spread over* its neighboring pixels.

Definition 8. A single bright pixel is called a point source and the function which describes the blurring process is called the point spread function (PSF). When the PSF is the same regardless of the location of the point source, we say that the blurring process is spatially invariant. Note that in spatially invariant blurring processes, the PSF contains all the information about the blurring.

Since it is natural to assume that the blurring process is a local phenomenon, PSFs can be described using a matrix with very small dimensions in a spatially invariant blurring. Finally, assuming that the imaging process captures all light, the pixel values in the PSF must sum to 1.

Example 10. The following PSF tells us that in the blurring process each pixel in the sharp image is replaced by the average of its nearest neighbors.

$$\boldsymbol{P} = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{4.1}$$

Having the PSF at our disposal, it should be clear how the matrix A is constructed. Due to the special structure of the matrix A, the matrix-vector product $A\boldsymbol{x}$ is called *convolution*. In Figure 4.2, we provide an example of a artificially blurred image using a 81×81 PSF and a white Gaussian noise.

It is commonly assumed that the sharp image $\boldsymbol{x}^*$ and the blurred noisy image $\boldsymbol{b}$ have equal dimensions. As a result, A can be considered as a square matrix. However, this assumption ignores the behavior of the blurring process at the boundary of the image. For example, consider the PSF defined in (4.1). How do we compute the corresponding blurred pixels which lie at boundary of the sharp image? One common approach called *padding with zeros* is used to address this situation wherein we extend our sharp image by assuming zero pixel values all the way around it. Since the value of these extended artificial pixels is already known, we do not need to incorporate them into the unknown vector $\boldsymbol{x}^*$. On the other hand, the resultant extra artificial pixels inside the blurry noisy image $\boldsymbol{b}$ are kept inside the noisy blurry image $\boldsymbol{b}$ as they provide useful information about the boundary of the sharp image. As a result, $A \in \boldsymbol{R}^{m\times n}$ is no longer a square matrix and instead is rectangular with $m \geq n$. Finally, the matrix A is supposed to be full column rank due to the way it is constructed (using some small PSF matrix).

Assumption 4. We always assume that the blurring matrix $A \in \boldsymbol{R}^{m\times n}$ has more rows than columns *i.e.*, $m > n$. Furthermore, we assume that A is full column rank meaning that $A^TA \in \boldsymbol{R}^{n\times n}$ is non-singular.

We conclude this section by emphasizing that in the rest of this chapter

$$\text{Assumptions 2, 3 and 4 hold,} \tag{4.2}$$

and also the following notation is adopted.

Notation 1. Denote the spectral decomposition of A^TA as follows:

$$A^TA = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T \text{ where } \lambda_1 \geq \cdots \geq \lambda_n > 0, \tag{4.3}$$

and the singular values of A by $\sigma_i := \sqrt{\lambda_i}$ for all $i = 1, \cdots, n$. Moreover, we express the components of $\boldsymbol{x}^*$ in the basis of $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n\}$ by $x_1^*, \cdots, x_n^*$, *i.e.*,

$$x_i^* := \langle \boldsymbol{x}^*, \boldsymbol{u}_i \rangle.$$

$\{x_i^*\}_{i=1}^n$ are called the singular value expansion (SVE) coefficients of $\boldsymbol{x}^*$ w.r.t. A.

4.1.1 The discrete Picard condition

It is well-known that as i increases the eigenvector $\boldsymbol{u}_i$ tends to have more sign changes. As a result, the spectral decomposition defined in (4.3) can be used for an expansion where each $\boldsymbol{u}_i$ represents a certain frequency. It is also well-known that, in this basis, most images are described by their dominated low-frequency spectral components as the high-frequency ones are smaller in magnitude (Book [34], Chapter 1, page 10). For future reference, we state this fact as follows.

Fact 8. In the context of image deblurring, a standard assumption is that the SVE coefficients $x_i^* = \langle \boldsymbol{x}^*, \boldsymbol{u}_i \rangle$ decay faster than the singular values σ_i. This condition is known as *discrete Picard condition*, see *e.g.* (3.16) in [33].

We emphasize that some stronger relationship between the decay rates of SVE coefficients and the singular values is often considered. These heuristic assumptions help to make

Figure 4.2: A sharp image (left) and its corresponding blurred image. The PSF $\frac{1}{81} \cdot \text{ones}(9, 9)$ is used for artificially blurring the sharp image. The noise level here equals to 0.05 *i.e.,* $\frac{\|\boldsymbol{\xi}\|}{\|A\boldsymbol{x}^*\|} \approx 0.015$ and $(m, n) = (10404, 10000)$.

the basic ideas and the techniques as clear as possible, rather than striving for maximal generality. One of such common assumptions is as follows.

$$|x_i^*| = \lambda_i^c \text{ for all } i = 1, \cdots, n, \tag{4.4}$$

where $c > 0$ is a fixed parameter, *e.g.* see Eq. (20) in [30].

The relationship between SVE coefficients and the singular values are illustrated using Picard plots. A detailed description of different varieties of PSFs, Picard plots and boundary conditions in deblurring are not discussed in detail in this section. We refer the reader to the books [34, 33].

4.2 Least square deconvolution

Let the blurring matrix A and the noisy blurred image $\boldsymbol{b}$ be mathematically modeled as above. The reconstruction of $\boldsymbol{x}^*$ from A and $\boldsymbol{b}$ is called deconvolution [32]. To this end, minimizing the norm of the residual $\|A\boldsymbol{x} - \boldsymbol{b}\|^2$ might seem a reasonable approach and thus we are led naturally to the following least squares problem.

$$\min \frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|^2. \tag{4.5}$$

Is the solution to (4.5) is a good estimate of the unknown vector $\boldsymbol{x}^*$? In view of Assumption 4, we can compute the solution to (4.5) explicitly:

$$\boldsymbol{x}_{\text{LS}} := \left(A^T A\right)^{-1} A^T \boldsymbol{b}. \tag{4.6}$$

Recall that $\boldsymbol{b} = A\boldsymbol{x}^* + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim N(\boldsymbol{0}, \sigma^2 I_m)$. Therefore, we can rewrite (4.6) to obtain:

$$\boldsymbol{x}_{\text{LS}} = \boldsymbol{x}^* + \underbrace{\left(A^T A\right)^{-1} A^T \boldsymbol{\xi}}_{:=\boldsymbol{\xi}_{\text{LS}}}.$$

We observe that $\boldsymbol{x}_{\text{LS}}$ is contaminated with the noise $\boldsymbol{\xi}_{\text{LS}}$. Using Fact 4, we can easily verify that

$$\boldsymbol{\xi}_{\text{LS}} \sim N(\boldsymbol{0}, \sigma^2 \left(A^T A\right)^{-1}). \tag{4.7}$$

By (4.3) and (4.7), we have that

$$\mathbb{E}\left[\|\boldsymbol{\xi}_{\text{LS}}\|^2\right] = \sigma^2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \tag{4.8}$$

where we used Facts 4 and 7. Combining (4.8) and the following fact reveals to us why $\boldsymbol{x}_{\text{LS}}$ is not a good estimate for the unknown vector $\boldsymbol{x}^*$.

Fact 9. For a blurring matrix A, the singular values decay gradually to zero. As such, the condition number of A which is defined as $\frac{\sigma_1}{\sigma_n}$ is very large. Recall that, in this situation, we say that matrix A is *ill-conditioned*.

All in all, the reason that $\boldsymbol{x}_{\text{LS}}$ cannot be used as a good estimate of $\boldsymbol{x}^*$ is due to the amplification of high-frequency components of the noise in the data and this is caused by the inversion of very small singular values of the ill-conditioned blurring matrix A. In the next section, we recall the commonly used regularization methods from Section 1.5 and the optimization algorithms used along these regularization method for reconstructing the sharp image $\boldsymbol{x}^*$ by solving (4.5).

4.2.1 Regularization

As previously reasoned in Section 4.2, solving the least squares problem (4.5) to optimality does not necessary yield a desirable estimate of the sharp image $\boldsymbol{x}^*$ and it may even lead to a worse blurring than $\boldsymbol{b}$. We also explained that the reason for this is that the inversion of the very small singular values of the blurring matrix A will amplify the high-frequency components of the noise in the data. We discussed in Section 1.5 that regularization methods are the most commonly used approach for dealing with this phenomenon. In particular, we stated the Tikhonov regularization method. We also mentioned that various regularization methods for gradient-based algorithms such as GD, LSMR, LSQR and other CG-type methods have been already researched. As far as we know, the previous literature on the least squares deconvolution did not explore the performance of the SGD algorithm for least squares deconvolution (Algorithm 9). Algorithm 9 will be further explored in the rest of this chapter and particularly, the following observations are made.

- It is illustrated through numerical and theoretical evidence that SGD with early stopping exhibits implicit regularization. Our theory in particular shows that SGD and GD shall exhibit similar behaviour.
- Inspired by numerical experiments, we introduce a new easily implementable and inexpensive stopping rule for SGD. Motivated by the fact that both GD and SGD behave similarly and that the GD algorithm is more amenable to analysis, we analyze our stopping time for the GD algorithm instead of SGD.

Algorithm 9: SGD algorithm for least squares deconvolution

initialize: $\boldsymbol{x}_0^{\text{SGD}} = 0$, A and $\boldsymbol{b}$ as in (4.2), $\alpha > 0$

for $k = 0, 1, 2, \cdots$

Choose $i_k \in \{1, 2, \cdots, m\}$ uniformly at random.

Update $\boldsymbol{x}_k^{\text{SGD}}$ by setting

$$\boldsymbol{x}_{k+1}^{\text{SGD}} = \boldsymbol{x}_k^{\text{SGD}} - \alpha \left(\left\langle A[i_k, :]^T, \boldsymbol{x}_k^{\text{SGD}} \right\rangle - \boldsymbol{b}_{i_k} \right) \cdot A[i_k, :]^T. \tag{4.9}$$

$k \leftarrow k + 1$

end

4.3 SGD with early stopping

We begin this section by computing the expected value of the error squared term for the GD algorithm (Algorithm 10).

Algorithm 10: GD algorithm for least squares deconvolution

initialize: $\boldsymbol{x}_0^{\text{GD}} = \mathbf{0} \in \boldsymbol{R}^n$, A and $\boldsymbol{b}$ as in (4.2), $\alpha > 0$

for $k = 0, 1, \cdots$

Update $\boldsymbol{x}_{k+1}^{\text{GD}} = \boldsymbol{x}_k^{\text{GD}} - \alpha A^T \left(A \boldsymbol{x}_k^{\text{GD}} - \boldsymbol{b} \right)$

$k \leftarrow k + 1$

end

Proposition 3. *Let $\{\boldsymbol{x}_k^{GD}\}_{k=0}^{+\infty}$ be the sequence generated by Algorithm 10. The following is then true for all $N \geq 0$.*

$$\mathbb{E}\left[\|\boldsymbol{x}_N^{GD} - \boldsymbol{x}^*\|^2 \right] = \sum_{i=1}^{n} \beta_i^{2N} (x_i^*)^2 + \sigma^2 \sum_{i=1}^{n} \frac{\left(1 - \beta_i^N\right)^2}{\lambda_i}, \tag{4.10}$$

where $\beta_i = 1 - \alpha\lambda_i$. The eigenvalues λ_i and the SVE coefficients x_i^* for $i = 1, \cdots, n$ are defined in Notation 1.

Proof. Update formula of GD can be written as follows.

$$\boldsymbol{x}_k^{\text{GD}} = \left(I - \alpha \cdot A^T A \right) \boldsymbol{x}_{k-1}^{\text{GD}} + \alpha A^T \boldsymbol{b}. \tag{4.11}$$

Iterating (4.11) and taking into account that $\boldsymbol{x}_0 = \mathbf{0}$, we obtain that

$$\begin{aligned}
\boldsymbol{x}_N^{\mathrm{GD}} &= \alpha \left[I + (I - \alpha A^T A) + \cdots + (I - \alpha A^T A)^{N-1} \right] A^T \boldsymbol{b} \\
&= \alpha \left[I + (I - \alpha A^T A) + \cdots + (I - \alpha A^T A)^{N-1} \right] A^T \left(A\boldsymbol{x}^* + \boldsymbol{\xi} \right) \\
&= \alpha \left[I + (I - \alpha A^T A) + \cdots + (I - \alpha A^T A)^{N-1} \right] \left(A^T A\boldsymbol{x}^* + A^T A\boldsymbol{\xi}_{\mathrm{LS}} \right) \\
&= \alpha \left[I + (I - \alpha A^T A) + \cdots + (I - \alpha A^T A)^{N-1} \right] A^T A \left(\boldsymbol{x}^* + \boldsymbol{\xi}_{\mathrm{LS}} \right) \\
&= \left(I - (I - \alpha A^T A)^N \right) \left(\boldsymbol{x}^* + \boldsymbol{\xi}_{\mathrm{LS}} \right)
\end{aligned} \tag{4.12}$$

In the third equality, we used that $A^T \boldsymbol{\xi} = A^T A \boldsymbol{\xi}_{\mathrm{LS}}$. By (4.12), we conclude that

$$\boldsymbol{x}_N^{\mathrm{GD}} - \boldsymbol{x}^* = -(I - \alpha A^T A)^N \boldsymbol{x}^* + \left(I - (I - \alpha A^T A)^N \right) \boldsymbol{\xi}_{\mathrm{LS}}. \tag{4.13}$$

Hence, it holds that

$$\begin{aligned}
\|\boldsymbol{x}_N^{\mathrm{GD}} - \boldsymbol{x}^*\|^2 = \|(I - \alpha A^T A)^N \boldsymbol{x}^*\|^2 + \| \left(I - (I - \alpha A^T A)^N \right) \boldsymbol{\xi}_{\mathrm{LS}} \|^2 \\
- 2\boldsymbol{\xi}_{\mathrm{LS}}^T \left(I - (I - \alpha A^T A)^N \right) (I - \alpha A^T A)^N \boldsymbol{x}^*.
\end{aligned} \tag{4.14}$$

By Fact 4 and (4.7), we can see that

$$\mathbb{E}\left[\boldsymbol{\xi}_{\mathrm{LS}}^T \left(I - (I - \alpha A^T A)^N \right) (I - \alpha A^T A)^N \boldsymbol{x}^* \right] = 0 \tag{4.15}$$

Furthermore by Fact 7 and (4.7), it can be verified that

$$\mathbb{E}\left[\| \left(I - (I - \alpha A^T A)^N \right) \boldsymbol{\xi}_{\mathrm{LS}} \|^2 \right] = \sigma^2 \operatorname{Tr}\left(\left(I - \left(I - \alpha A^T A \right)^N \right) \left(A^T A \right)^{-1} \right). \tag{4.16}$$

Combining (4.14), (4.15) and (4.16), (4.10) readily follows. □

We next illustrate that the implicit regularization effect of GD with early stopping for least squares deconvolution follows from Facts 8, 9 and Proposition 3 under the assumption that

$$\alpha \leq \frac{1}{\lambda_1}. \tag{4.17}$$

To this end, in view of Fact 9, let us suppose that for some $1 \leq s \leq n$ the following approximation hold.

$$\lambda_i \approx 0 \quad \text{for all } i \geq s. \tag{4.18}$$

By Fact 8 and (4.18), we obtain that

$$x_i^* \approx 0 \quad \text{for all } i \geq s. \tag{4.19}$$

By (4.18), (4.19) and the fact that β_i for $1 \leq i \leq s$ converges to 0 rapidly, we are led to believe that the following hold for some small k.

$$\sum_{i=1}^{n} \beta_i^{2k} (x_i^*)^2 \approx 0 \text{ and } \sigma^2 \sum_{i=1}^{n} \frac{\left(1-\beta_i^k\right)^2}{\lambda_i} \approx \sigma^2 \sum_{i=1}^{s} \frac{1}{\lambda_i}. \tag{4.20}$$

Notice that we used the assumption (4.17). On the other hand, we observe that

$$\mathbb{E}\left[\|\boldsymbol{x}_{\mathrm{LS}} - \boldsymbol{x}^*\|^2\right] = \sigma^2 \sum_{i=1}^{n} \frac{1}{\lambda_i}. \tag{4.21}$$

Finally, combining (4.18), (4.20) and (4.21), we get that

$$\mathbb{E}\left[\|\boldsymbol{x}_{T_{\mathrm{GD}}}^{\mathrm{GD}} - \boldsymbol{x}^*\|^2\right] \ll \mathbb{E}\left[\|\boldsymbol{x}_{\mathrm{LS}} - \boldsymbol{x}^*\|^2\right]. \tag{4.22}$$

Hence, if GD is halted at the optimal stopping time, it will exhibit an implicit regularization effect in the sense of (4.22).

4.3.1 Implicit regularization of SGD with early stopping

We already reasoned that GD with early stopping for least squares deconvolution exhibits favorable regularization. This section is devoted to understanding the implicit regularization effect of SGD (Algorithm 9) with early stopping.

The following Theorem shows that the iterates generated by the SGD algorithm with sufficiently small step-size follow the gradient flow with overwhelming probability.

Theorem 9. *Let $\{\boldsymbol{x}_k^{SGD}\}_{k=0}^{+\infty}$ be a sequence generated by Algorithm 9. Fix $i \in \{1, \cdots, n\}$ and let $\boldsymbol{u}_i$ be an eigenvector of $A^T A$ as in Notation 1. Moreover, suppose that $\alpha\lambda_i < m$. The following concentration bound then holds for all $N \geq 0$ and all positive real t.*

$$\mathbb{P}\left(\left|\left\langle \boldsymbol{u}_i, \boldsymbol{x}_N^{SGD} - \left(1 - \left(1 - \frac{\alpha\lambda_i}{m}\right)^N\right) \boldsymbol{x}_{LS} \right\rangle\right| \geq \frac{t}{\sigma_i}\right) \leq 2\exp\left(-\frac{t^2}{2\alpha m \tau_N^2}\right), \tag{4.23}$$

where τ_N^2 is a variance parameter depending on A, $\boldsymbol{b}$ and N. Particularly, it holds that $\tau_1 \leq \cdots \leq \tau_N$.

Notice that, ideally, it must hold that $m\tau_N = \mathcal{O}(1)$. In a discussion after the proof of Theorem 9, we illustrate that under some reasonable assumption this bound holds.

Furthermore, by (4.12), $\left(1-\left(1-\frac{\alpha\lambda_i}{m}\right)^N\right)\boldsymbol{x}_{\mathrm{LS}}$ equals to $\boldsymbol{x}_N^{\mathrm{GD}}$ where the GD sequence is generated using the step-size $\frac{\alpha}{m}$. Finally, note that

$$1-\left(1-\frac{\alpha\lambda_i}{m}\right)^{Nm}\approx 1-\exp\left(-\alpha\lambda_i N\right)\approx 1-\left(1-\alpha\lambda_i\right)^N.$$

Thus, the SGD and GD iterates indeed lie close to each other in the sense of Theorem 9.

Proof of Theorem 9. Update formula (4.9) can be rewritten as follows.

$$\begin{aligned}\boldsymbol{x}_{k+1}^{\mathrm{SGD}}&=\left(I-\alpha A[i_k,:]^TA[i_k,:]\right)\boldsymbol{x}_k^{\mathrm{SGD}}+\alpha\boldsymbol{b}_{i_k}A[i_k,:]^T\\&=\left(I-\alpha A[i_k,:]^TA[i_k,:]\right)\boldsymbol{x}_k^{\mathrm{SGD}}+\alpha A[i_k,:]^TA[i_k,:]\boldsymbol{x}^*+\alpha\boldsymbol{\xi}_{i_k}A[i_k,:]^T.\end{aligned}\tag{4.24}$$

Therefore, we conclude that

$$\boldsymbol{x}_{k+1}^{\mathrm{SGD}}-\boldsymbol{x}^*=\left(I-\alpha A[i_k,:]^TA[i_k,:]\right)\left(\boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}^*\right)+\alpha\boldsymbol{\xi}_{i_k}A[i_k,:]^T.\tag{4.25}$$

Taking conditional expectations from both sides in (4.25) verifies that for all $k\geq 1$ it holds.

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{x}_{k+1}^{\mathrm{SGD}}-\boldsymbol{x}^*|i_0,\cdots,i_{k-1}\right]&=\left(I-\frac{\alpha}{m}\cdot A^TA\right)\left(\boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}^*\right)+\frac{\alpha}{m}A^T\boldsymbol{\xi}\\&=\left(I-\frac{\alpha}{m}\cdot A^TA\right)\left(\boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}^*\right)+\frac{\alpha}{m}A^TA\boldsymbol{\xi}_{\mathrm{LS}}.\end{aligned}\tag{4.26}$$

Hence, by (4.26), we obtain that

$$\mathbb{E}\left[\boldsymbol{x}_{k+1}^{\mathrm{SGD}}-\boldsymbol{x}_{\mathrm{LS}}|i_0,\cdots,i_{k-1}\right]=\left(I-\frac{\alpha}{m}\cdot A^TA\right)\left(\boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}_{\mathrm{LS}}\right).\tag{4.27}$$

Denote by $z_k:=\langle\boldsymbol{u}_i,\boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}_{\mathrm{LS}}\rangle$. Taking inner products from both sides of (4.27) with $\boldsymbol{u}_i$, we see that

$$\mathbb{E}\left[z_{k+1}|i_0,\cdots,i_{k-1}\right]=\left(1-\frac{\alpha\lambda_i}{m}\right)z_k.\tag{4.28}$$

Denote by $\tilde{z}_k:=\left(1-\frac{\alpha\lambda_i}{m}\right)^{-k}z_k$. By (4.28), we have that

$$\mathbb{E}\left[\tilde{z}_{k+1}|i_0,\cdots,i_{k-1}\right]=\tilde{z}_k.$$

We next upper bound the difference $|\tilde{z}_{k+1}-\tilde{z}_k|$. To this end, using (4.24), it can be seen that there exists some positive constant M such that the following bound holds for all $k\geq 0$.

$$\left|\langle\boldsymbol{u}_i,\boldsymbol{x}_{k+1}^{\mathrm{SGD}}-\boldsymbol{x}_k^{\mathrm{SGD}}\rangle\right|\leq\alpha M.$$

Notice that

$$\begin{aligned}\left(1-\frac{\alpha\lambda_i}{m}\right)^{k+1}\cdot|\tilde{z}_{k+1}-\tilde{z}_k| &= \left|\langle \boldsymbol{u}_i, \boldsymbol{x}_{k+1}^{\mathrm{SGD}}-\boldsymbol{x}_{\mathrm{LS}}\rangle-\left(1-\frac{\alpha\lambda_i}{m}\right)\langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}_{\mathrm{LS}}\rangle\right| \\ &= \left|\langle \boldsymbol{u}_i, \boldsymbol{x}_{k+1}^{\mathrm{SGD}}-\boldsymbol{x}_k^{\mathrm{SGD}}\rangle+\frac{\alpha\lambda_i}{m}\langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}_{\mathrm{LS}}\rangle\right| \\ &\le \alpha M+\frac{\alpha\lambda_i}{m}\left|\langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}_{\mathrm{LS}}\rangle\right| \\ &\le \alpha\left(M+\frac{\lambda_i c_k}{m}\right),\end{aligned} \tag{4.29}$$

where c_k is defined by

$$c_k := \sup\left|\langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\mathrm{SGD}}-\boldsymbol{x}_{\mathrm{LS}}\rangle\right| < +\infty.$$

Here the supremum is taken over possible values of $\boldsymbol{x}_k$ for every possible choice of $i_0, \cdots, i_k$ which is finite set and hence the supremum is finite. Set

$$\tau_N := \sup_{k=1,\cdots,N} M+\frac{\lambda_i c_k}{m}.$$

The bound in (4.29) gives

$$\left(1-\frac{\alpha\lambda_i}{m}\right)^{k+1}|\tilde{z}_{k+1}-\tilde{z}_k| \le \alpha\tau_N.$$

By Azuma's inequality (Lemma 4) the following concentration bound holds for all positive real t.

$$\mathbb{P}\left(|\tilde{z}_N-\tilde{z}_0| \ge t\left(1-\frac{\alpha\lambda_i}{m}\right)^{-N}\right) \le 2\exp\left(-\frac{t^2}{2\tau_N^2\alpha^2\sum_{k=1}^N\left(1-\frac{\alpha\lambda_i}{m}\right)^{2(N-k)}}\right).$$

To obtain (4.23), it suffices to verify the following bound.

$$\alpha^2\sum_{k=1}^N\left(1-\frac{\alpha\lambda_i}{m}\right)^{2(N-k)} \le \frac{m\alpha}{\lambda_i}.$$

This bound obviously holds and hence (4.23) holds. The proof is complete. □

Ideally, it must hold that $m\tau_N^2 = \mathcal{O}(\frac{1}{m})$. To this end, we need to make the following assumption.

$$\sigma_i \left|\langle \boldsymbol{u}_i, \boldsymbol{x}_k^{\text{SGD}} - \boldsymbol{x}_{\text{LS}} \rangle\right| = \mathcal{O}(\sigma) + \mathcal{O}(1). \tag{4.30}$$

It is worth noting that the sequence $\{\boldsymbol{x}_k^{\text{SGD}}\}_{k=0}^{+\infty}$ converges to a neighborhood of $\boldsymbol{x}_{\text{LS}}$ [70]. As a result, (4.30) follows since $\sigma_i \langle \boldsymbol{u}_i, \boldsymbol{x}_{\text{LS}} \rangle \sim N(0, \sigma^2)$. Now, under the assumption that (4.30) holds, to obtain $m\tau_N^2 = \mathcal{O}(\frac{1}{m})$, it suffices to have that

$$\frac{m}{\alpha} \left|\langle \boldsymbol{u}_i, \boldsymbol{x}_{k+1}^{\text{SGD}} - \boldsymbol{x}_k^{\text{SGD}} \rangle\right| = \mathcal{O}(1). \tag{4.31}$$

Rewriting (4.31), we obtain that

$$m \left|\left(\langle A[i_k, :]^T, \boldsymbol{x}_k^{\text{SGD}} \rangle - \boldsymbol{b}_{i_k}\right) \cdot \langle \boldsymbol{u}_i, A[i_k, :]^T \rangle\right| = \mathcal{O}(1). \tag{4.32}$$

By the design of the blurring matrix A, we expect to have

$$\left|\langle A[i_k, :]^T, \boldsymbol{x}_k^{\text{SGD}} \rangle - \boldsymbol{b}_{i_k}\right| = \mathcal{O}(1) \text{ and } \left|\langle \boldsymbol{u}_i, A[i_k, :]^T \rangle\right| = \mathcal{O}(\frac{1}{m}). \tag{4.33}$$

Combining (4.30), (4.32) and (4.33), we obtain that $m\tau_N^2 = \mathcal{O}(\frac{1}{m})$. We leave it for later work to provide a rigorous proof for (4.30).

4.4 Numerical experiments

We conduct experiments where three algorithms are considered: SGD (Algorithm 9), GD (Algorithm 5) and CGLS (Algorithm 2). It is emphasized that a full experimental comparison of the SGD, GD, and CGLS algorithms for the deconvolution least squares problem is outside of the scope of this thesis; our numerical experiments is only a motivation for studying the SGD algorithm.

We denote $\{\boldsymbol{x}_k^{\text{SGD}}\}_{k=0}^{+\infty}$, $\{\boldsymbol{x}_k^{\text{GD}}\}_{k=0}^{+\infty}$ and $\{\boldsymbol{x}_k^{\text{CGLS}}\}_{k=0}^{+\infty}$ as the iterates of SGD, GD and CGLS respectively. For each algorithm, we record the number of iterations it takes for the error term $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$ to achieve its minimum value (denoted by T_{SGD}, T_{GD} and T_{CGLS} for SGD, GD and CGLS respectively). For example,

$$T_{\text{GD}} := \underset{k \geq 0}{\text{argmin}} \, \|\boldsymbol{x}_k^{\text{GD}} - \boldsymbol{x}^*\|.$$

We also calculate the relative error at the instant when the iteration count is reached. For the case of the GD algorithm, we can define

$$E_{\text{GD}} := \frac{\|\boldsymbol{x}_{T_{\text{GD}}}^{\text{GD}} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^*\|}, \tag{4.34}$$

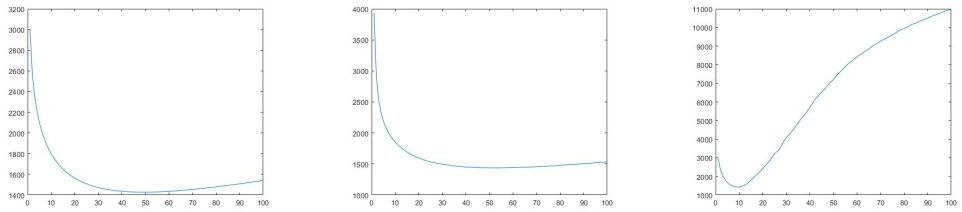

Figure 4.3: The error terms $\|\boldsymbol{x}_k^{\mathrm{GD}} - \boldsymbol{x}^*\|$ (left), $\|\boldsymbol{x}_k^{\mathrm{SGD}} - \boldsymbol{x}^*\|$ (right), and $\|\boldsymbol{x}_k^{\mathrm{CGLS}} - \boldsymbol{x}^*\|$ (center) where $\{\boldsymbol{x}_k^{\mathrm{SGD}}\}_{k=0}^{+\infty}$, $\{\boldsymbol{x}_k^{\mathrm{GD}}\}_{k=0}^{+\infty}$ and $\{\boldsymbol{x}_k^{\mathrm{CGLS}}\}_{k=0}^{+\infty}$ are the iterates of SGD, GD and CGLS algorithms applied to the least-squares problem (4.5) respectively. Here $\boldsymbol{x}^*$, A and $\boldsymbol{b}$ are constructed as in Figure 4.2. We observe that $T_{\mathrm{SGD}}, T_{\mathrm{GD}}$ and T_{CGLS} are equal to $54m$, 48 and 9 respectively and also $E_{\mathrm{SGD}} \approx E_{\mathrm{GD}} \approx E_{\mathrm{CGLS}}$.

where E_{GD} is the relative error. E_{SGD} and E_{CGLS} are the corresponding relative errors for SGD and CGLS algorithms receptively and are defined similarly to (4.34).

We perform a numerical experiment (Figure 4.3) to compute the values of $T_{\mathrm{SGD}}, T_{\mathrm{GD}}$, T_{CGLS} , $E_{\mathrm{SGD}}, E_{\mathrm{GD}}$ and E_{CGLS}. We observe that if the three algorithms are halted at their respective optimal stopping time (namely T_{SGD}, T_{GD} and T_{CGLS}), then they will produce approximately equal relative errors. In other words, we have

$$E_{\mathrm{SGD}} \approx E_{\mathrm{GD}} \approx E_{\mathrm{CGLS}}. \tag{4.35}$$

Furthermore, we have

$$T_{\mathrm{SGD}} \approx mT_{\mathrm{GD}}. \tag{4.36}$$

Moreover, in the event where the noise level is high enough, we observed that

$$T_{\mathrm{SGD}} < mT_{\mathrm{GD}}. \tag{4.37}$$

It is emphasized that across several more experiments, we observed similar results as in (4.35), (4.36) and (4.37). In our experiments, we plotted the sequences $\left\{\|A\boldsymbol{x}_k^{\mathrm{SGD}}\|\right\}_{k=0}^{+\infty}$ and observed that for the GD and SGD algorithms both plots are Γ-shaped (Figure 4.4). Particularly, as the noise level increases, the corner of the Γ-curve becomes closer to the value $\|A\boldsymbol{x}_{T_{\mathrm{SGD}}}^{\mathrm{SGD}}\|$. On the other hand, in a low noise regime, the error term at termination is more robust to any unsubstantial miscalculation of the optimal stopping time. Motivated by these observations, we propose the following stopping rule for the SGD algorithm:

$$T_{\mathrm{S}} := \inf\{k : \|A\boldsymbol{x}_{km}^{\mathrm{SGD}}\|^2 \geq \delta\|\boldsymbol{b}\|^2\}, \tag{4.38}$$

where $\delta \in (0, 1)$ is a hyperparameter. In the next section, we provide theoretical analysis for the stopping rule (4.38).

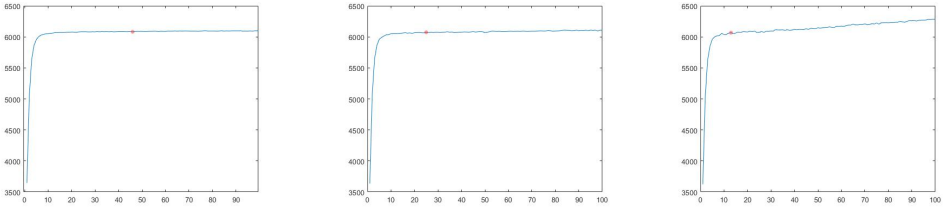

Figure 4.4: Plots of the sequence $\{\|A\boldsymbol{x}_k^{\mathrm{SGD}}\|\}_{k=0}^{+\infty}$ where the sequence $\{\boldsymbol{x}_k^{\mathrm{SGD}}\}_{k=0}^{+\infty}$ is generated by Algorithm 9. From left to right, the noise levels are equal to $0.015, 0.03$ and 0.06 respectively. The corresponding value of $\|A\boldsymbol{x}_{T_{\mathrm{SGD}}}^{\mathrm{SGD}}\|$ is plotted by a red dot.

4.5 Stopping time analysis

As a first step towards understanding the behaviour of the stopping time (4.38), we provide an analysis for the case where the iterates of the GD algorithm (Algorithm 11) are used instead of SGD.

$$T := \inf\left\{k : \|A\boldsymbol{x}_k^{\mathrm{GD}}\|^2 \geq \delta\|\boldsymbol{b}\|^2\right\}. \tag{4.39}$$

Algorithm 11: GD for LS deconvolution with stopping criterion

initialize: $\boldsymbol{x}_0^{\mathrm{GD}} = \boldsymbol{0} \in \boldsymbol{R}^n$, A and $\boldsymbol{b}$ as in (4.2), $\alpha > 0$, $\delta \in (0,1)$
while $\|A\boldsymbol{x}_k\|^2 < \delta\|\boldsymbol{b}\|^2$
  Update $\boldsymbol{x}_{k+1}^{\mathrm{GD}} = \boldsymbol{x}_k^{\mathrm{GD}} - \alpha A^T\left(A\boldsymbol{x}_k^{\mathrm{GD}} - \boldsymbol{b}\right)$
  $k \leftarrow k+1$
set $T = k$
end

We defer the analysis of (4.38) for later work (Section 5.2). The next theorem provides a theoretical upper bound for the ℓ_2-error term at termination.

Theorem 10. *Let $\{\boldsymbol{x}_k^{GD}\}_{k=0}^{+\infty}$, T and δ be as in Algorithm 11. Assume that for some $r \in \{1, \cdots, n\}$ the following bound holds where $\{\lambda_i\}_{i=1}^n$ and $\{x_i^*\}_{i=1}^n$ are defined in Notation 1.*

$$\sum_{i>r} \lambda_i \left(x_i^*\right)^2 \leq \frac{1-\delta}{2} \sum_{i\leq r} \lambda_i \left(x_i^*\right)^2. \tag{4.40}$$

Suppose that $\sigma \leq \frac{1-\delta}{4} \cdot \frac{\|\sqrt{A^TA}\boldsymbol{x}^*\|}{\sqrt{m+1}}$ and α satisfies $\alpha\lambda_1 \leq \frac{1}{2}$. Then with probability at least

$1-2\exp\left(-\frac{m+1}{8}\right)$, *$T$ is finite and the following bound holds.*

$$\|\boldsymbol{x}_T^{GD}-\boldsymbol{x}^*\| \leq \sqrt{\sum_{i=1}^{n}\exp\left(2\log\left(4\left(1-\delta\right)\right)\frac{\lambda_i}{\lambda_1}\right)x_i^{*2}} + \sigma\sqrt{\sum_{i=1}^{n}\min\left\{\frac{1}{\lambda_i},\log\left(\frac{1-\delta}{4}\right)^2\frac{\lambda_i}{\lambda_r^2}\right\}}.$$

Here $\boldsymbol{x}_T^{GD}$ denotes the iterate at termination.

Before turning to the proof of Theorem 10, it is important to note that the bound (4.40) is guaranteed by assuming that the magnitude of the SVE coefficients x_i^* decay faster than the eigenvalues λ_i. Ideally, when using the assumption (4.4), the value of $c>0$ should be large enough. We already explained in 4.1.1 that in the context of image deblurring, such fast decay rates are prevalent (Fact 8 and bound (4.4)). A large enough value of c and by extension a fast decay rate will also ensure that

$$\frac{\lambda_1}{\lambda_r} \text{ is not large for } r \text{ in (4.40).} \tag{4.41}$$

Under these heuristic assumptions, for properly chosen δ, we have

$$\sum_{i=1}^{n}\exp\left(2\log\left(4\left(1-\delta\right)\right)\frac{\lambda_i}{\lambda_1}\right)x_i^{*2} \ll \|\boldsymbol{x}^*\|^2 \;\&\; \sum_{i=1}^{n}\min\left\{\frac{1}{\lambda_i},\log\left(\frac{1-\delta}{4}\right)^2\frac{\lambda_i}{\lambda_r^2}\right\} \ll \sum_{i=1}^{n}\frac{1}{\lambda_i}.$$

Figure 4.5 exemplifies the fact that the assumed decay rate for the SVE coefficients is appropriate in such a way that the informal assumption (4.41) also holds.

Ideally, Theorem 10 can be used for providing a good stopping criterion for GD if the parameter δ is properly tuned. However, it is emphasized that Theorem 10 is only for the purpose of providing insight for the SGD least squares deconvolution algorithm, hence estimating the best value of δ is not a concern.

Before proving Theorem 9, we need three auxiliary lemmas.

Lemma 14. *Let A,α and δ be as in Algorithm 11. Denote by*

$$\Phi_k := \left(I-\left(I-\alpha A^TA\right)^k\right)^2-\delta I.$$

Suppose that $\alpha<\frac{1}{\lambda_1}$. The following then holds.

$$\Phi_0 \preceq \Phi_1 \preceq \Phi_2 \preceq \cdots.$$

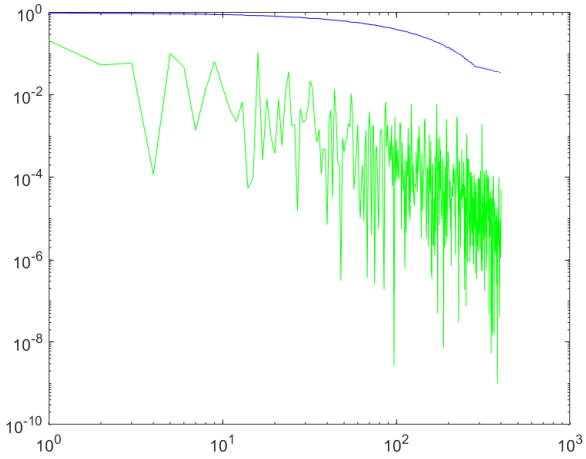

Figure 4.5: Plot of the decay rates for $\{\log(\lambda_i)\}_{i=1}^n$ (blue curve) and $\{\log((x_i^*)^2)\}_{i=1}^n$ (green curve). Here the same data as in Figure 4.2 is used. It can be observed that for $r = 500$, the informal assumption (4.41) holds.

Proof. Note that $\Phi_k \succeq 0$ for all $k \geq 0$. Fix $k \geq 0$. We observe that

$$\begin{aligned}
\Phi_k \preceq \Phi_{k+1} &\iff \left(I - \left(I - \alpha A^T A\right)^k\right)^2 \preceq \left(I - \left(I - \alpha A^T A\right)^{k+1}\right)^2 \\
&\iff I - \left(I - \alpha A^T A\right)^k \preceq I - \left(I - \alpha A^T A\right)^{k+1} \\
&\iff \left(I - \alpha A^T A\right)^{k+1} \preceq \left(I - \alpha A^T A\right)^k.
\end{aligned}$$

The second implication holds by using the simple fact from elementary linear algebra stating that for two commuting positive semidefinite matrices X and Y, it holds that $X \preceq Y$ if and only if $X^2 \preceq Y^2$. □

Lemma 15. *Let $\alpha, w_1, \cdots, w_r, x_1, \cdots, x_r \in (0, +\infty)$ such that $w_r \leq \cdots \leq w_1$ and $\alpha w_1 \leq \frac{1}{2}$. Fix $\ell, u \in (0, 1)$ and assume that positive real t satisfies the following bound.*

$$\ell \leq \frac{\sum_{i=1}^r \exp\left(t \log(1 - \alpha w_i)\right) w_i x_i}{\sum_{i=1}^r w_i x_i} \leq u. \tag{4.42}$$

The following is then true.

$$\frac{-\log(u)}{2w_1} \leq t\alpha \leq \frac{-\log(\ell)}{w_r}. \tag{4.43}$$

Proof. We begin by denoting $\mu_i = \log(1-\alpha w_i)$ and $p_i = \frac{w_i x_i}{\sum_{i=1}^r w_i x_i}$ for all $i \in [r]$[2]. Note that $p_1, \cdots, p_r$ define a probability distribution on $[r]$. Therefore, for any positive real t satisfying (4.42), it holds that

$$\ell \leq \mathbb{E}\left[\exp\left(t\mu_i\right)\right] \leq u. \tag{4.44}$$

Using Jensen's inequality, we obtain that

$$\exp\left(t\mathbb{E}\left[\mu_i\right]\right) \leq \mathbb{E}\left[\exp\left(t\mu_i\right)\right]$$

By (4.44) and (4.76), we obtain that

$$t\sum_{i=1}^r \mu_i p_i \leq \log(u) \tag{4.45}$$

Applying the inequality $\log(1+b) \geq \frac{b}{1+b}$ for all $b > -1$, we then obtain that

$$-\log(1-\alpha w_i) = -\mu_i \leq \frac{\alpha w_i}{1-\alpha w_i} \leq 2\alpha w_i, \tag{4.46}$$

where the assumption $\frac{1}{1-\alpha w_i} \leq 2$ for all $i \in [r]$ is used. Combining (4.45) and (4.46), we have that

$$-\log\left(u\right) \leq 2t\alpha \sum_{i=1}^r w_i p_i. \tag{4.47}$$

By (4.47) and $\sum_{i=1}^n p_i = 1$, we have

$$-\log(u) \leq 2t\alpha w_1.$$

We thus obtain the LHS inequality in (4.43). To obtain the other bound in (4.43), since $\mu_1 \leq \cdots \leq \mu_r$, we have that

$$\ell \leq \mathbb{E}\left[\exp\left(t\mu_i\right)\right] \leq \exp(t\mu_r) \leq \exp(-t\alpha w_r), \tag{4.48}$$

where the bound $\log(1-b) \leq -b$ for all $b < 1$ is used. By (4.48), we see that

$$t\alpha \leq \frac{-\log(\ell)}{w_r}$$

The proof is complete. □

[2] $[r] := \{1, \cdots, r\}$

Lemma 16. *Let $\{\boldsymbol{x}_k^{GD}\}_{k=0}^{+\infty}$, α and δ be as in Algorithm 11. The following identity holds.*

$$\|A\boldsymbol{x}_N^{GD}\|^2 - \delta\|\boldsymbol{b}\|^2 = (\boldsymbol{y}^* + \boldsymbol{\psi})^T \Phi_N (\boldsymbol{y}^* + \boldsymbol{\psi}) - \delta\|\boldsymbol{\psi}_0\|^2, \tag{4.49}$$

where Φ_k is defined in (14), $\boldsymbol{y}^* := \sqrt{A^TA}\boldsymbol{x}^*$, $\boldsymbol{\psi} := \left(A^TA\right)^{\frac{1}{2}} \boldsymbol{\xi}_{LS}$ and $\boldsymbol{\psi}_0 \sim N(0, \sigma^2 I_{m-n})$. Particularly, the sequence $\{\|A\boldsymbol{x}_k^{GD}\|\}_{k=0}^{+\infty}$ is monotonically increasing provided that $\alpha < \frac{1}{\lambda_1}$.

Proof. By (4.12), it holds that

$$\boldsymbol{x}_N^{\mathrm{GD}} = \left(I - (I - \alpha A^TA)^N\right)(\boldsymbol{x}^* + \boldsymbol{\xi}_{\mathrm{LS}}). \tag{4.50}$$

This yields that

$$\begin{aligned}
\|A\boldsymbol{x}_N^{\mathrm{GD}}\|^2 &= \left(\boldsymbol{x}_N^{\mathrm{GD}}\right)^T A^TA\boldsymbol{x}_N^{\mathrm{GD}} \\
&= (\boldsymbol{x}^* + \boldsymbol{\xi}_{\mathrm{LS}})^T \left(I - \left(I - \alpha A^TA\right)^N\right) A^TA \left(I - \left(I - \alpha A^TA\right)^N\right)(\boldsymbol{x}^* + \boldsymbol{\xi}_{\mathrm{LS}}) \\
&= (\boldsymbol{x}^* + \boldsymbol{\xi}_{\mathrm{LS}})^T \sqrt{A^TA} \left(I - \left(I - \alpha A^TA\right)^N\right) \left(I - \left(I - \alpha A^TA\right)^N\right) \sqrt{A^TA} (\boldsymbol{x}^* + \boldsymbol{\xi}_{\mathrm{LS}}) \\
&= (\boldsymbol{y}^* + \boldsymbol{\psi})^T (\Phi_N + \delta I)(\boldsymbol{y}^* + \boldsymbol{\psi})
\end{aligned}$$

In addition,

$$\begin{aligned}
\|\boldsymbol{b}\|^2 = (\boldsymbol{x}^*)^T A^TA\boldsymbol{x}^* + 2(\boldsymbol{x}^*)^T A^T\boldsymbol{\xi} + \|\boldsymbol{\xi}\|^2 &= \|\boldsymbol{y}^*\|^2 + 2(\boldsymbol{y}^*)^T\boldsymbol{\psi} + \|\boldsymbol{\xi}\|^2 \\
&= (\boldsymbol{y}^* + \boldsymbol{\psi})^T (\boldsymbol{y}^* + \boldsymbol{\psi}) + \left(\|\boldsymbol{\xi}\|^2 - \|\boldsymbol{\psi}\|^2\right).
\end{aligned}$$

Combining the pieces, we see that

$$\|A\boldsymbol{x}_N^{\mathrm{GD}}\|^2 - \delta\|\boldsymbol{b}\|^2 = (\boldsymbol{y}^* + \boldsymbol{\psi})^T \Phi_N (\boldsymbol{y}^* + \boldsymbol{\psi}) - \delta\left(\|\boldsymbol{\xi}\|^2 - \|\boldsymbol{\psi}\|^2\right). \tag{4.51}$$

Using $\boldsymbol{\psi} = \left(A^TA\right)^{-\frac{1}{2}} A^T\boldsymbol{\xi}$, it holds that

$$\|\boldsymbol{\xi}\|^2 - \|\boldsymbol{\psi}\|^2 = \boldsymbol{\xi}^T\boldsymbol{\xi} - \boldsymbol{\psi}^T\boldsymbol{\psi} = \boldsymbol{\xi}^T\boldsymbol{\xi} - \boldsymbol{\xi}^T A \left(A^TA\right)^{-1} A^T\boldsymbol{\xi}. \tag{4.52}$$

Let $I - A\left(A^TA\right)^{-1}A^T = WD_0W^T$ where $W^TW = I_m$ and D_0 is a $m \times m$ diagonal matrix with the first $m-n$ diagonal entries equal to 1 and the rest equal to 0. Letting $\boldsymbol{\psi}_0$ to be the first $m-n$ entries of $D_0W^T\boldsymbol{\xi}$ we obtain (4.49). Combining (4.49) and Lemma 14, we obtain that $\|A\boldsymbol{x}_k^{\mathrm{GD}}\|$ is increasing. The proof is complete. □

Next proposition introduces deterministic stopping times τ_1 and τ_2 and shows that T (Eq. (4.39)) lies in $[\tau_1, \tau_2]$ with overwhelming probability provided that σ (Assumption 3) is small enough. The proof is based on the concentration of measure for norm of random Gaussian vectors.

Proposition 4. *Define deterministic values τ_1 and τ_2 as follows.*

$$\tau_1 := \max\left\{k : (\boldsymbol{y}^*)^T \Phi_k \boldsymbol{y}^* \leq -(1-\delta)\cdot\|\boldsymbol{y}^*\|^2\right\} \quad \textit{and} \quad \tau_2 := \min\left\{k : (\boldsymbol{y}^*)^T \Phi_k \boldsymbol{y}^* \geq (1-\delta)\cdot\|\boldsymbol{y}^*\|^2\right\}.$$

Here Φ_k is defined in Lemma 14 and $\boldsymbol{y}^* := \sqrt{A^TA}\boldsymbol{x}^*$. Provided that $\sigma \leq \frac{(1-\delta)}{8}\cdot\frac{\|\boldsymbol{y}^*\|}{\sqrt{m+1}}$, then with probability at least $1-2\exp\left(-\frac{m+1}{8}\right)$, it holds that $T \in [\tau_1, \tau_2]$.

Proof. By Lemma 5, the following holds with probability at least $1-2\exp\left(-\frac{m+1}{8}\right)$.

$$\|\boldsymbol{\psi}_0\| \leq \sigma(\tfrac{1}{2}\sqrt{m-n+1}+\epsilon) \leq \sigma\cdot\sqrt{m+1}, \quad \|\boldsymbol{\psi}\| \leq \sigma(\tfrac{1}{2}\sqrt{n+1}+\epsilon) \leq \sigma\sqrt{m+1}, \tag{4.53}$$

where we set $\epsilon = \frac{1}{2}\sqrt{m+1}$. By Lemma 16, $\|A\boldsymbol{x}_k^{\mathrm{GD}}\|$ is monotonic for $k = 0, 1, 2, \cdots$. Therefore, T is finite if and only if the following bound holds.

$$\|A\boldsymbol{x}_{LS}\|^2 - \delta\|\boldsymbol{b}\|^2 = (1-\delta)\|\boldsymbol{y}^* + \boldsymbol{\psi}\|^2 - \delta\|\boldsymbol{\psi}_0\|^2 \geq 0. \tag{4.54}$$

Assuming that the bounds in (4.53) hold, we obtain that

$$\begin{aligned}
\sqrt{1-\delta}\cdot\|\boldsymbol{y}^*+\boldsymbol{\psi}\| &\geq \sqrt{1-\delta}\cdot\|\boldsymbol{y}^*\| - \sqrt{1-\delta}\cdot\|\boldsymbol{\psi}\| \\
&\geq \sqrt{1-\delta}\cdot\|\boldsymbol{y}^*\| - \sqrt{1-\delta}\cdot\sigma\cdot\sqrt{m+1} \\
&\geq \sqrt{1-\delta}\cdot\frac{8\sigma}{1-\delta}\cdot\sqrt{m+1} - \sqrt{1-\delta}\cdot\sigma\cdot\sqrt{m+1} \\
&\geq \sigma\cdot\sqrt{m+1}\cdot\left(\frac{8}{\sqrt{1-\delta}} - \sqrt{1-\delta}\right) \\
&\geq \sigma\cdot\sqrt{m+1} \\
&\geq \|\boldsymbol{\psi}_0\|.
\end{aligned}$$

We used the bound $\sigma \leq \frac{1-\delta}{4}\cdot\frac{\|\boldsymbol{y}^*\|}{\sqrt{m+1}}$ in the third inequality. Therefore T is finite assuming that the bound (4.53) hold. We then have that[3]

$$(\boldsymbol{y}^*+\boldsymbol{\psi})^T\,\Phi_T\,(\boldsymbol{y}^*+\boldsymbol{\psi}) \approx \delta\|\psi_0\|^2 \tag{4.55}$$

[3]It is clear that there exists $t^* \in [T-1, T]$ such that $(\boldsymbol{y}^*+\boldsymbol{\psi})^T\,\Phi_{t^*}\,(\boldsymbol{y}^*+\boldsymbol{\psi}) = \delta\|\psi_0\|^2$. Because of this, the bounds in Theorem 9 for the error term at t^* instead of T. However, for simplicity of expression, we assume that this approximation holds.

By (4.53), we have

$$\begin{aligned}\left|\delta\|\boldsymbol{\psi}_0\|^2 - \boldsymbol{\psi}^T\Phi_k\boldsymbol{\psi} - 2\boldsymbol{\psi}^T\Phi_k\boldsymbol{y}^*\right| &\le \sigma^2\left(\epsilon + \tfrac{1}{2}\sqrt{m-n+1}\right)^2 + \sigma^2\left(\epsilon + \tfrac{1}{2}\sqrt{n+1}\right)^2 \\ &\quad + 2\sigma\|\boldsymbol{y}^*\|\left(\epsilon + \tfrac{1}{2}\sqrt{n+1}\right) \\ &\le 2\sigma^2(m+1) + 2\sigma\|\boldsymbol{y}^*\|\sqrt{m+1} \\ &\le 2\left(\tfrac{1}{16}(1-\delta)^2 + \tfrac{1}{4}(1-\delta)\right)\|\boldsymbol{y}^*\|^2 \\ &\le (1-\delta)\|\boldsymbol{y}^*\|^2.\end{aligned}$$

Here we used the bounds $\sigma\sqrt{m+1} \le \frac{\|\boldsymbol{y}^*\|}{2}$. Therefore, with probability at least $1 - 2\exp\left(-\frac{m+1}{8}\right)$, the following bound holds for all k.

$$\left|\delta\|\boldsymbol{\psi}_0\|^2 - \boldsymbol{\psi}^T\Phi_k\boldsymbol{\psi} - 2\boldsymbol{\psi}^T\Phi_k\boldsymbol{y}^*\right| \le (1-\delta)\|\boldsymbol{y}^*\|^2 \tag{4.56}$$

By (4.55) and (4.56), we obtain that the following estimate holds with probability at least $1 - 2\exp\left(-\frac{m+1}{8}\right)$,

$$\left|(\boldsymbol{y}^*)^T\Phi_T\boldsymbol{y}^*\right| \le (1-\delta)\|\boldsymbol{y}^*\|^2.$$

This immediately implies that $T \in [\tau_1, \tau_2]$ with probability at least $1 - 2\exp\left(-\frac{m+1}{8}\right)$. The proof is complete. □

We are now ready to prove Theorem 9.

Proof of Theorem 9. Instantiate notation from Proposition 4. Denote $\mu_i := \log(1-\alpha\lambda_i)$ and recall the deterministic stopping times τ_1 and τ_2. Rearranging yields that

$$[\tau_1, \tau_2] \subseteq \left\{t : 1 - \delta \le \frac{1}{\|\boldsymbol{y}^*\|^2}\sum_{i=1}^{n}\left(1 - (1-\exp(t\mu_i))^2\right)\lambda_i x_i^{*2} \le 2(1-\delta)\right\}. \tag{4.57}$$

We will first show that

$$[\tau_1, \tau_2] \subseteq \left\{t : -\log(4(1-\delta)) \cdot \frac{1}{\lambda_1} \le t\alpha \le -\log\left(\frac{1-\delta}{4}\right) \cdot \frac{1}{\lambda_r}\right\}. \tag{4.58}$$

Using $\exp(t\mu_i) \le 1 - (1-\exp(t\mu_i))^2 \le 2\exp(t\mu_i)$, we obtain that

$$[\tau_1, \tau_2] \subseteq \left\{t : \frac{1-\delta}{2} \le \frac{1}{\|\boldsymbol{y}^*\|^2}\sum_{i=1}^{n}\exp(t\mu_i)\lambda_i x_i^{*2} \le 2(1-\delta)\right\}. \tag{4.59}$$

By Assumption (4.40), we have

$$\frac{1}{\|\boldsymbol{y}^*\|^2}\sum_{i\geq r+1}\lambda_i x_i^{*2}\leq\frac{1-\delta}{2}. \tag{4.60}$$

By (4.59) and (4.60), we have

$$[\tau_1,\tau_2]\subseteq\left\{t:\frac{1-\delta}{4}\leq\frac{1}{\|\boldsymbol{y}^*\|^2}\sum_{i=1}^{r}\exp(t\mu_i)\lambda_i x_i^{*2}\leq 2(1-\delta)\right\}. \tag{4.61}$$

Using Assumption (4.40), we obtain that $\sum_{i\geq r+1}\lambda_i x_i^{*2}\leq\sum_{i=1}^{r}\lambda_i x_i^{*2}$. Therefore,

$$\frac{\sum_{i=1}^{r}\exp(t\mu_i)\lambda_i x_i^{*2}}{2\sum_{i=1}^{r}\lambda_i x_i^{*2}}\leq\frac{1}{\|\boldsymbol{y}^*\|^2}\sum_{i=1}^{r}\exp(t\mu_i)\lambda_i x_i^{*2}\leq\frac{\sum_{i=1}^{r}\exp(t\mu_i)\lambda_i x_i^{*2}}{\sum_{i=1}^{r}\lambda_i x_i^{*2}}. \tag{4.62}$$

By (4.59), (4.61) and (4.62), we obtain that

$$[\tau_1,\tau_2]\subseteq\left\{\frac{1-\delta}{4}\leq\frac{\sum_{i=1}^{r}\exp(t\mu_i)\lambda_i x_i^{*2}}{\sum_{i=1}^{r}\lambda_i x_i^{*2}}\leq 4(1-\delta)\right\}. \tag{4.63}$$

By Lemma 15, the bound on α ($\alpha\lambda_1\leq\frac{1}{2}$) and (4.63), we obtain (4.58) as desired. Next, denote

$$C_b:=\sup_{t\geq\tau_1}\sqrt{\sum_{i=1}\exp\left(2t\mu_i\right)x_i^{*2}},\quad\text{and}\quad C_v:=\frac{\sigma}{\sigma_i}\cdot\sup_{0\leq t\leq\tau_2}|1-\exp(t\mu_i)|. \tag{4.64}$$

Our next aim is to upper bound C_b and C_v. Let $k\in[\tau_1,\tau_2]$. By (4.58), we have that

$$-\log\left(3(1-\delta)\right)\cdot\frac{1}{\lambda_1}\leq\tau_1\leq k \tag{4.65}$$

By (4.65), we have

$$\exp\left(2k\mu_i\right)\leq\exp\left(-2k\alpha\lambda_i\right)\leq\exp\left(2\log\left(3\left(1-\delta\right)\right)\cdot\frac{\lambda_i}{\lambda_1}\right)\quad\forall i\in[n], \tag{4.66}$$

In order to bound C_v, using the inequality $1+x\leq\exp(x)$, we obtain that

$$\frac{\sigma}{\sigma_i}\cdot(1-\exp(k\mu_i))\leq\frac{\sigma}{\sigma_i}\cdot(-k\mu_i)=\sigma\sigma_i\cdot k\alpha\cdot\frac{\log(1-\alpha\lambda_i)}{-\alpha\lambda_i}\leq\sigma\sigma_i\cdot k\alpha. \tag{4.67}$$

We also clearly have that

$$\frac{\sigma}{\sigma_i} \cdot \sup_{0 \le t \le \tau_2} |1 - \exp(t\mu_i)| \le \frac{\sigma}{\sigma_i}. \tag{4.68}$$

By (4.58), (4.67) and (4.68), we obtain that

$$C_v^2 \le \sigma^2 \sum_{i=1}^{n} \min \left\{ \frac{1}{\lambda_i}, \log\left(\frac{1-\delta}{4}\right)^2 \cdot \frac{\lambda_i}{\lambda_r^2} \right\}. \tag{4.69}$$

The proof is complete. □

4.6 A matrix concentration inequality for products

In this section, we present a *non-asymptotic concentration inequality* for the random matrix product

$$Z_n = (I_d - \alpha X_n)(I_d - \alpha X_{n-1}) \cdots (I_d - \alpha X_1), \tag{4.70}$$

where $\{X_k\}_{k=1}^{+\infty}$ is a sequence of bounded independent random positive semidefinite matrices with common expectation $\mathbb{E}[X_k] = \Sigma$. Under these assumptions, we show that, for small enough positive α, Z_n satisfies the concentration inequality

$$\mathbb{P}(\|Z_n - \mathbb{E}[Z_n]\| \ge t) \le 2d^2 \cdot \exp\left(\frac{-t^2}{\alpha\sigma^2}\right) \quad \text{for all } t \ge 0, \tag{4.71}$$

where σ^2 denotes a variance parameter. We remark that the bound (4.71) suggests that for the least-squared problem the SGD algorithm trajectory follows the gradient flow with overwhelming probability. See Figure 4.6.

Products of random matrices appear as building blocks for many stochastic iterative algorithms, *e.g.* [67, 81]. While non-asymptotic bounds of averages of these matrices are well developed, *e.g.* [83, 86], the analogous bounds of their products are much harder to understand due to the non-commutative nature of matrix multiplication. As such, efforts to understand bounds of this type have become an active area of research *e.g.* [39, 43, 47]. In this note, we provide a non-asymptotic concentration bound (4.71) for the random matrix product Z_n (4.70). These instances appear, for example, in the stochastic gradient descent algorithm applied to the linear least squares problem. We remark that bound (4.71) will be of special interest when X_k is almost surely low rank for all k. In this event, almost all eigenvalues of each factor in the matrix product Z_n are equal to 1 whereas $\mathbb{E}[Z_n]$ has an exponentially decaying operator norm. (Note: Without loss of generality, we can assume

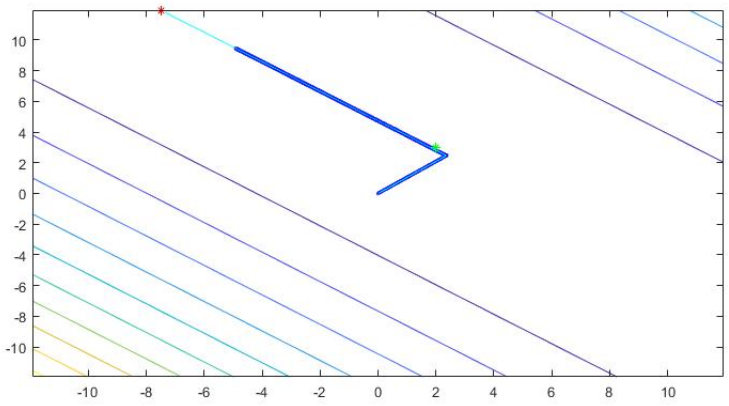

Figure 4.6: We consider the SGD algorithm applied to the least-squares problem with $A \in \boldsymbol{R}^{3\times 2}$, $\sigma = 0.5$ and $\alpha = 0.001$. The diagonal lines are the level sets of the objective function, the green asterisk is $\boldsymbol{x}^*$, the red asterisk is $\boldsymbol{x}_{LS}$, the path of blue dots are the SGD iterates, and the light-blue curve is the gradient flow.

that Σ is positive definite.) Hence, it is interesting to observe that Z_n concentrates around its mean with overwhelming probability as in (4.71), especially in the case where X_k's are almost surely low rank matrices.

In [43], using the uniform smoothness property of the Schatten p-norm, the authors have studied non-asymptotic bounds for the products of random matrices, in particular, random contractions [43, Theorem 7.1]. To apply their result to the matrix product (4.70), we will need to make some further assumptions. First, we need to assume some bound involving $|I - \alpha X_k|$ since the Araki-Lieb-Thirring inequality [8, IX.2.11] is used in their analysis. Second, we need to assume a lower bound $t^2 \geq c\alpha^2 \sum_{k=1}^{n} \|X_k - \Sigma\|^2$ which may grow linearly in n. This will be problematic particularly since we are only interested in the case where $t \leq 1$.

On the other hand, compared to our result, the bound in [43, Theorem 7.1] has a weaker dependency on the dimension d and, more importantly, it works in a broader variety of instances. For example, one can use their bound when in (4.70), instead of $I - \alpha X_k$, we consider the factors $I - \alpha_k X_k$ with α_k decaying at a proper rate.

We next provide a proof for (4.71). The proof proceeds by constructing a martingale sequence satisfying bounded differences and then applying Azuma's inequality, Lemma 4. We assume that the positive semidefinite random matrices X_k in (4.70) are drawn independently and they satisfy $\mathbb{E}[X_k] = \Sigma$ for all k. In addition, we suppose that X_k are

uniformly bounded in the operator norm, meaning that there exists $r > 0$ such that

$$\|X_k\|_{\text{op}} \leq r \quad \text{almost surely.}$$

Let $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_d$ denote the eigenvectors of Σ and $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ denote the corresponding eigenvalues. For each $i \in [d] := \{1, \cdots, d\}$, define c_i to be the infimum over all positive reals for which

$$\|(X_k - \lambda_i I)\, \boldsymbol{\mu}_i\|_{\text{op}} \leq c_i \lambda_i \quad \text{almost surely.}$$

Note that $c_i < +\infty$ almost surely as $c_i \leq 1 + \frac{r}{\lambda_i}$ and also, because X_k is positive semidefinite, $c_i = 0$ whenever $\lambda_i = 0$. We will use the following parameter to measure the amount of variation in X_k

$$\sigma^2 := \frac{4d}{3} \sum_{i=1}^{d} c_i^2 \lambda_i.$$

Theorem 11. *Suppose that $\alpha \in (0, \frac{1}{2r})$. Then the following concentration inequality holds.*

$$\mathbb{P}\left(\|Z_n - \mathbb{E}\left[Z_n\right]\|_{op} \geq t \right) \leq 2d^2 \cdot \exp\left(\frac{-t^2}{\alpha \sigma^2} \right). \tag{4.72}$$

Proof. Without loss of generality, we can assume that $c_i, \lambda_i > 0$ for all $i \in [d]$. We will first show that, for any $i, j \in [d]$, the following bound holds for all $t \geq 0$:

$$\mathbb{P}\left(\left| \boldsymbol{\mu}_i^T Z_n \boldsymbol{\mu}_j - \mathbb{E}\left[\boldsymbol{\mu}_i^T Z_n \boldsymbol{\mu}_j \right] \right| \geq t \sqrt{\tfrac{4\lambda_i}{3}} \cdot c_i \right) \leq 2 \exp\left(\frac{-t^2}{\alpha} \right). \tag{4.73}$$

Set $Z_0 = I_d$. Then we note that for all $k \geq 0$,

$$\mathbb{E}\left[\boldsymbol{\mu}_i^T Z_k \boldsymbol{\mu}_j \right] = \boldsymbol{\mu}_i^T \mathbb{E}\left[Z_k\right] \boldsymbol{\mu}_j = \boldsymbol{\mu}_i^T \left(I - \alpha \Sigma\right)^k \boldsymbol{\mu}_j = \left(1 - \alpha \lambda_i\right)^k \cdot \delta_{i,j},$$

where $\delta_{i,j}$ stands for Kronecker delta. For notational convenience, let us denote $z_k := \boldsymbol{\mu}_i^T Z_k \boldsymbol{\mu}_j$. We have that

$$\mathbb{E}\left[z_k | X_{k-1}, \cdots, X_1 \right] = \boldsymbol{\mu}_i^T \mathbb{E}\left[I - \alpha X_k\right] Z_{k-1} \boldsymbol{\mu}_j = \left(1 - \alpha \lambda_i\right) z_{k-1}. \tag{4.74}$$

Denote $q_i := 1 - \alpha \lambda_i$ and define the random variable $Y_k := q_i^{-k} \cdot z_k$. Dividing both sides of (4.74) by q_i^k, we obtain that $\mathbb{E}\left[Y_k | X_{k-1}, \cdots, X_1\right] = Y_{k-1}$. Thus, $\{Y_k\}_{k=1}^{+\infty}$ is a martingale with respect to $\{X_k\}_{k=1}^{+\infty}$. We observe that for all $k \geq 1$

$$q_i^k \cdot |Y_k - Y_{k-1}| = |z_k - q_i \cdot z_{k-1}| = \alpha \left| \boldsymbol{\mu}_i^T \left(X_k - \lambda_i I\right) Z_{k-1} \boldsymbol{\mu}_j \right| \leq \alpha c_i \lambda_i,$$

where the assumption $\alpha r \leq \frac{1}{2}$ yielded the bound $\|Z_{k-1}\| \leq 1$ a.s. Thus, by Azuma's inequality Lemma 4, we have that for any $\epsilon \geq 0$

$$\begin{aligned}\mathbb{P}\left(|z_n - \mathbb{E}\left[z_n\right]| \geq \epsilon\right) &= \mathbb{P}\left(|Y_n - Y_0| \geq \epsilon \cdot q_i^{-n}\right) \\ &\leq 2\exp\left(\frac{-\epsilon^2}{2\alpha^2\lambda_i^2 c_i^2 \sum_{k=0}^{n-1} q_i^{2k}}\right).\end{aligned} \tag{4.75}$$

Note that by Jensen's inequality

$$\lambda_i \leq \|\Sigma\| = \|_{\text{op}}\mathbb{E}\left[X_k\right]\|_{\text{op}} \leq \mathbb{E}\left[\|X_k\|_{\text{op}}\right] \leq r. \tag{4.76}$$

Therefore, by (4.76) and since $\alpha \in (0, \frac{1}{2r})$, we obtain that $\sum_{k=0}^{n-1} q_i^{2k} \leq \frac{2}{3(1-q_i)}$. Plugging this bound into the right-hand side of (4.75) and letting $\epsilon = t\sqrt{\frac{4\lambda_i}{3}} \cdot c_i$, we will obtain (4.73). Finally, in order to see (4.72), we observe that by (4.73), with probability exceeding $1 - 2d^2 \cdot \exp\left(-\frac{t^2}{\alpha}\right)$, it holds that

$$\|Z_n - \mathbb{E}\left[Z_n\right]\|^2 \leq t^2 \cdot \frac{4d}{3}\sum_{i=1}^{d} c_i^2\lambda_i.$$

Therefore,

$$\mathbb{P}\left(\|Z_n - \mathbb{E}\left[Z_n\right]\| \geq t \cdot \sigma\right) \leq 2d^2 \cdot \exp\left(\frac{-t^2}{\alpha}\right).$$

The result immediately follows since $\|Z_n - \mathbb{E}\left[Z_n\right]\|_{\text{op}} \leq \|Z_n - \mathbb{E}\left[Z_n\right]\|$. □

Chapter 5

Conclusion and Future Work

While first order methods induce very low computational cost and require low memory storage, they still need to be regularized to prevent overfitting. The most classical type of regularization is by way of adding a penalty function to the objective function. An alternative form however is based on a method called early stopping, in which we halt the algorithm once some termination criterion has been activated. Imposing much less computational cost is the main advantage of early stopping over other forms of regularization.

Because of its various favorable aspects, the stochastic gradient descent algorithm and its variants play a key role in modern optimization. Scalability for large scale models [37] and parallelizability with big training data [28] are among the most important features of the SGD algorithm. Yet, despite these benefits, simple easily implemented termination criteria for SGD are not well-studied. In this thesis, we studied the SGD algorithm with a termination criterion for two fundamental problems: binary classification and the least squares.

5.1 Key results

The key results of this thesis are summarized below.

I We have proposed a simple and computationally free termination test for SGD for binary classification, supported by both theoretical and experimental results. The theoretical results show that the test will stop SGD after a finite time with a bound on the expected accuracy of the resulting classifier. In our experimental results, the plots

show consistent pattern that our test achieves low accuracy but is faster than SVS for $\tilde{\alpha} = 1/10$, while it achieves higher accuracy with more iterations when $\tilde{\alpha} = 1/200$. This is useful behavior in practice, compared to SVS, since it puts the accuracy/iterations trade-off in the hands of the user who selects the step-size $\tilde{\alpha}$. Another benefit of our new termination criterion apparent from all plots is that the number of iterations is more consistent across random trials, which is beneficial in the case that SGD is used as a sub-problem of a larger computation. racy for the iteration at termination.

II In the case of least-squares problem, we considered the deconvolution task using the SGD algorithm. We established a novel concentration bound to show that for a small enough step-size, the SGD path shall follow the gradient flow trajectory. Motivated by numerical observation, we proposed a new termination criterion for the SGD algorithm for the least squares deconvolution problem. As a first step towards developing theoretical guarantees for our termination criterion, we provide an upper bound for the ℓ_2-error term for the iterate at termination when the gradient descent algorithm is considered.

5.2 Future work

This thesis has raised many questions, some remained unanswered:

1. It will be interesting if we extend our results in Chapter 3, especially Theorem 6, to the case where the logistic or hinge loss functions are replaced by some non-convex function *e.g.,* quasi-convex functions.

2. We would like to extend our results in the high variance regime (Theorem 7) in Chapter 3. We need to establish an upper bound for $\mathbb{E}[T]$ based on the data parameters.

3. As observed in our numerical experiments, our test exhibits high accuracy on a broader range of distributions. Therefore, extending our results in Chapter 3 to a more general class of distributions rather than Gaussian would be interesting.

4. In Chapter 3, we studied the binary *linear* classification problem. Using kernel methods, our results should be extendable to non-linear cases as well. We leave this for later work.

5. Matrix product in the form of (4.70) appears ubiquitously in optimization. For example, using the results in [27], we observe that at least six different well-known

algorithms for solving consistent linear systems including the randomized Kaczmarz method, randomized Newton method, randomized coordinate descent method and random Gaussian pursuit have the following update rule, [27] Eq. 2.8,

$$\boldsymbol{x}_n - \boldsymbol{x}^* = (I - X_n) \cdots (I - X_1) (\boldsymbol{x}_0 - \boldsymbol{x}^*) , \tag{5.1}$$

where X_k is a random matrix drawn i.i.d from some fixed distribution $\mathcal{P}$. Therefore, using results in [27], we can express the iterates of many randomized optimization algorithms in the form of (5.1) where the corresponding probability distribution is expressed in a unified manner. This motivates the following problem: Can we extend results of [2] to obtain similar concentration bounds for the update rules (5.1) where $X_k \sim \mathcal{P}$ i.i.d and $\mathcal{P}$ is a given distribution as in [27]? In particular, can we use these new bounds to improve the worst-case running time of algorithms for solving linear systems? Recently in [44], the authors have used concentration of inequalities to study the convergence property of streaming k-PCA.

6. Establishing theoretical upper bounds for the ℓ_2-error term at termination for the following stopping time.

$$T_{\mathrm{S}} := \inf \left\{ k : \|A\boldsymbol{x}_k^{\mathrm{SGD}}\|^2 \geq \delta \|\boldsymbol{b}\|^2 \right\} .$$

Bibliography

[1] Noga Alon and Joel H Spencer. *The Probabilistic Method*. John Wiley & Sons, 2004.

[2] Sina Baghal. A matrix concentration inequality for products, 2020.

[3] Sina Baghal, Courtney Paquette, and Stephen A. Vavasis. A termination criterion for stochastic gradient descent for binary classification, 2020.

[4] Johnathan M Bardsley. Applications of a nonnegatively constrained iterative method with statistically based stopping rules to CT, PET, and SPECT imaging. *Electronic Transactions on Numerical Analysis*, 38:34–43, 2011.

[5] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2006.

[6] Mario Bertero and Patrizia Boccacci. *Introduction to inverse problems in imaging*. CRC press, 2020.

[7] Dimitri P Bertsekas. *Convex Optimization Theory*. Athena Scientific Belmont, 2009.

[8] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

[9] J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

[10] Léon Bottou. Online Learning and Stochastic Approximations. *On-line Learning in Neural Networks*, 17(9):142, 1998.

[11] Léon Bottou and Olivier Bousquet. The Tradeoffs of Large-Scale Learning. *Optimization for Machine Learning*, page 351, 2011.

[12] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *Siam Review*, 60(2):223–311, 2018.

[13] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[14] S. Bubeck. *Convex Optimization: Algorithms and Complexity*, volume 8. Now Publishers Inc., 2015.

[15] Jerry Chee and Panos Toulis. Convergence diagnostics for stochastic gradient descent with constant learning rate. In *International Conference on Artificial Intelligence and Statistics*, pages 1476–1485. PMLR, 2018.

[16] D. Drusvyatskiy and D. Davis. Robust stochastic optimization with the proximal point method. *preprint arXiv:1907.13307*, 2019.

[17] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[18] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, New York, NY, USA, 4th edition, 2010.

[19] Tommy Elfving, Per Christian Hansen, and Touraj Nikazad. Convergence analysis for column-action methods in image reconstruction. *Numerical Algorithms*, 74(3):905–924, 2017.

[20] Tommy Elfving, Touraj Nikazad, and Per Christian Hansen. Semi-convergence and relaxation parameters for a class of SIRT algorithms. *Electronic Transactions on Numerical Analysis*, 37(274):321–336, 2010.

[21] F. Famoye. Continuous Univariate Distributions, volume 1. *Technometrics*, 37:466–466, 11 1995.

[22] David Chin-Lung Fong and Michael Saunders. LSMR: An Iterative Algorithm for Sparse Least-Squares Problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011.

[23] Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *Journal of Mathematics and Physics*, 38(1-4):77–81, 1959.

[24] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. *SIAM J. Optim.*, 22(4):1469–1492, 2012.

[25] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms. *SIAM J. Optim.*, 23(4):2061–2089, 2013.

[26] Gene H Golub and Urs Von Matt. Generalized cross-validation for large-scale problems. *Journal of Computational and Graphical Statistics*, 6(1):1–34, 1997.

[27] Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[28] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training Imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[29] Oleg Grodzevich and Henry Wolkowicz. Regularization using a parameterized trust region subproblem. *Mathematical Programming*, 116(1):193–220, 2009.

[30] Per Christian Hansen. Analysis of Discrete Ill-posed Problems by Means of the L-Curve. *SIAM review*, 34(4):561–580, 1992.

[31] Per Christian Hansen. The L-curve and its use in the numerical treatment of inverse problems. *Citeseer*, 1999.

[32] Per Christian Hansen. Deconvolution and Regularization with Toeplitz Matrices. *Numerical Algorithms*, 29(4):323–378, 2002.

[33] Per Christian Hansen. *Discrete Inverse Problems: Insight and Algorithms*. SIAM, 2010.

[34] Per Christian Hansen, James G Nagy, and Dianne P O'leary. *Deblurring Images: Matrices, Spectra, and Filtering*. SIAM, 2006.

[35] Per Christian Hansen, Victor Pereyra, and Godela Scherer. *Least squares data fitting with applications*. JHU Press, 2013.

[36] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[38] Jeff Heaton. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning, 2018.

[39] Amelia Henriksen and Rachel Ward. Concentration inequalities for random matrix products. *Linear Algebra and its Applications*, 594:81–94, 2020.

[40] Gabor T Herman. *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*. Springer Science & Business Media, 2009.

[41] Magnus Rudolph Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *NBS Washington, DC*, 49(1), 1952.

[42] David S Holder. *Electrical Impedance Tomography: Methods, History and Applications*. CRC Press, 2004.

[43] De Huang, Jonathan Niles-Weed, Joel A Tropp, and Rachel Ward. Matrix concentration for products. *arXiv preprint arXiv:2003.05437*, 2020.

[44] De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-PCA: Efficient guarantees for Oja's algorithm, beyond rank-one updates. *arXiv preprint arXiv:2102.03646*, 2021.

[45] Jean Jacod and Philip Protter. *Probability Essentials*. Springer Science & Business Media, 2012.

[46] Peter A Jansson. *Deconvolution of Images and Spectra*. Courier Corporation, 2014.

[47] Tarun Kathuria, Satyaki Mukherjee, and Nikhil Srivastava. On concentration inequalities for random matrix products. *arXiv preprint arXiv:2003.06319*, 2020.

[48] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[49] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[50] Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

[51] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[52] J. Lin, R. Camoriano, and L. Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2340–2348, 2016.

[53] Patrice Marcotte and Gilles Savard. Novel approaches to the discrimination problem. *Zeitschrift für Operations Research*, 36(6):517–545, 1992.

[54] Geoffrey J McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

[55] Paul D McNicholas. *Mixture Model-Based Classification*. CRC press, 2016.

[56] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

[57] D. Molitor, D. Needell, and R. Ward. Bias of homotopic gradient descent for the hinge loss. *preprint arXiv:1907.11746*, 2019.

[58] M. Nacson, N. Srebro, and D. Soudry. Stochastic Gradient Descent on Separable Data. In *Conference on Artificial Intelligence and Statistics*, 2019.

[59] Frank Natterer. *The mathematics of computerized tomography*. SIAM, 2001.

[60] Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

[61] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

[62] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[63] Arkadi Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Chichester: Wiley, 1983.

[64] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[65] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[66] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

[67] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

[68] Christopher C Paige and Michael A Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.

[69] Vivak Patel. Stopping criteria for, and strong convergence of, stochastic gradient descent on Bottou-Curtis-Nocedal functions, 2020.

[70] G. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM J. Control Optim.*, 24(4):655–666, 1986.

[71] David L Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM (JACM)*, 9(1):84–97, 1962.

[72] L. Prechelt. *Early Stopping — But When?*, pages 53–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[73] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

[74] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[75] D. A. Reynolds and R. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3, 02 1995.

[76] James D Riley. Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix. *Mathematical Tables and Other Aids to Computation*, pages 96–101, 1955.

[77] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[78] R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[79] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[80] Frank Spitzer. *Principles of Random Walk*, volume 34. Springer Science & Business Media, 2013.

[81] Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, 2009.

[82] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[83] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[84] E. van den Berg and M.P. Friedlander. Sparse optimization with least-squares constraints. *SIAM J. Optim.*, 21(4):1201–1229, 2011.

[85] A. van der Sluis and H.A. van der Vorst. SIRT- and CG-type methods for the iterative solution of sparse linear least-squares problems. *Linear Algebra and its Applications*, 130:257–303, 1990.

[86] Martin J Wainwright. *High-dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[87] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[88] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[89] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the 21st International Conference on Machine Learning*, page 116, 2004.

[90] J. Ziwei and M. Telgarsky. Risk and parameter convergence of logistic regression. *preprint arXiv:1803.07300*, 2018.