A termination criterion for stochastic gradient descent for binary classification

Sina Baghal SREZAZAD@UWATERLOO.CA
Courtney Paquette YUMIKO88@UW.EDU
Stephen Vavasis VAVASIS@UWATERLOO.CA
University of Waterloo, Canada

Abstract

We propose a new, simple, and computationally inexpensive termination test for constant step-size stochastic gradient descent (SGD) applied to binary classification on the logistic and hinge loss with homogeneous linear predictors. Our theoretical results support the effectiveness of our stopping criterion when the data is Gaussian distributed. This presence of noise allows for the possibility of non-separable data. 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. Finally, numerical experiments indicate for both real and synthetic data sets that our termination test exhibits a good degree of predictability on accuracy and running time.

1. Introduction

Minimization of an expected loss objective function using linear predictors,

$$\min_{\boldsymbol{\theta} \in \boldsymbol{R}^d} f(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{\zeta}, y) \sim \mathcal{P}} \ell(\boldsymbol{\zeta}^T \boldsymbol{\theta}, y), \tag{1}$$

is a central task in machine learning. Here the loss function is $\ell : \boldsymbol{R}^d \times \boldsymbol{R} \to \boldsymbol{R}$, the probability distribution $\mathcal{P}$ is unknown, and the data sample $(\boldsymbol{\zeta}, y) \in \boldsymbol{R}^d \times \boldsymbol{R}$ is a random vector distributed as $\mathcal{P}$. The most prevalent algorithm employed for solving (1) is *stochastic gradient descent* (SGD). Whereas a significant amount of work has been devoted to the convergence analysis of SGD (see, *e.g.*, [3, 4, 15, 16]), leading, in particular, to learning rate schedules, the question of how to terminate the algorithm when one is near an optimal classifier remains largely unaddressed.

Yet, inexpensive stopping criteria are of utmost interest in machine learning. For instance, if one could produce a low cost test to determine near-optimality, then without sacrificing the quality of the solution or efficiency of the SGD algorithm, needless computational time would be eliminated. Secondly, early termination tests impose a degree of predictability on accuracy and running times– a useful quality when SGD occurs as a subproblem of a larger computation. Several works show that early termination of SGD can prevent overfitting, speed up learning procedures, and/or improve generalization properties [7, 10, 21]. Motivated by these facts, we sought to address from stochastic optimization the following question:

How to design a test to terminate SGD with a fixed learning rate that is inexpensive without sacrificing quality of the solution?

To do so, we restrict ourselves to binary classification, one of the fundamental examples of supervised machine learning [19]. In binary classification, the learning algorithm is given a sequence of training examples $(\boldsymbol{\zeta}_1, y_1), (\boldsymbol{\zeta}_2, y_2), \ldots$, often noisy, where $\boldsymbol{\zeta}_i \in \boldsymbol{R}^d$ and $y_i \in \{0, 1\}$ for each i. The job of the algorithm is to develop a rule for distinguishing future, unseen $\boldsymbol{\zeta}$'s that are classified as 1 from those classified as 0. In this work, we limit our attention to linear classifiers. This means that the learning algorithm must determine a vector $\boldsymbol{\theta}$ such that the classification of $\boldsymbol{\zeta}$ is 1 when $\boldsymbol{\zeta}^T\boldsymbol{\theta} > 0$ else it is 0. Note that any algorithm for linear classification can be extended to one for nonlinear classification via the construction of "kernels"; see, *e.g.*, [19]. This extension is left for future work.

The usual technique for determining $\boldsymbol{\theta}$, which is also adopted here, is to define a loss function that turns the discrete problem of computing a 1 or 0 for $\boldsymbol{\zeta}$ to a continuous problem. Common choices of loss functions include logistic and hinge. For brevity, here we only consider unregularized logistic loss and we defer the analysis of hinge loss to Appendix.

Our contributions. In this paper, we introduce a new and simple termination criterion for stochastic gradient descent (SGD) applied to binary classification using logistic and hinge regression with constant step-size $\alpha > 0$. 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 (to be made precise later), we have the following results:

1. SGD with our stopping criterion will terminate for any fixed positive step-size. In particular, we establish an upper bound for the expected number of iterations before termination 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 exponentially 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. 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 [11] and CIFAR-10 [9]). In our experiments, we observe that our test yields relatively accurate classifiers with small variation across multiple runs.

Related work. The relationship between generalization and optimization is an active area of research in machine learning. Much of the pioneering work in this area focused on understanding how early termination of algorithms, such as conjugate gradient, gradient descent, and SGD, can act as an implicit regularizer and thus exhibit better generalization properties [10, 12, 13, 20, 21]. Most notably, to the best of our knowledge, the earliest comprehensive numerical testing of a stopping termination test for SGD in neural networks was introduced by [10]. His stopping criterion, which we denote as *small validation set* (SVS), periodically checks the iterate on a validation set. Theoretical guarantees for SVS were established in the works of [13, 21]. [7] shows that SGD is uniformly stable and thus solutions with low training error found quickly generalize well. These results support exploring new computationally inexpensive termination tests– the spirit of this paper.

2. Notation

Throughout the paper we consider a Euclidean space, denoted by $\boldsymbol{R}^d$, with an inner product and an induced norm $\|\cdot\|$. Bold-faced variables are vectors. The matrix I_d is the $d \times d$ identity matrix. The *indicator of the event* A is denoted by 1_A. If X is a measurable function and $t \in \boldsymbol{R}$, we often simplify the notation for the pull back of the function X, to simply $\{\omega \in \Omega \,:\, X(\omega) \leq t\} =: \{X \leq t\}$. Finally, in the analysis of our stopping criteria in Section 4, we borrow the notion of *stopping times* from probability theory. We refer the reader to [6] for related details.

3. Stopping criterion for stochastic gradient descent

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

$$\mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}}[\ell(\boldsymbol{\zeta}^T\boldsymbol{\theta}, y)] := \mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}}[-y\boldsymbol{\zeta}^T\boldsymbol{\theta} + \log(1+\exp(\boldsymbol{\zeta}^T\boldsymbol{\theta}))].$$

Here the samples $(\boldsymbol{\zeta}, y) \in \boldsymbol{R}^d \times \{0,1\}$. The data comes from a mixture model, that is, flip a fair coin to determine whether an item is in the $y=0$ or $y=1$ class, then generate the sample $\boldsymbol{\zeta}$ from either the distribution $\mathcal{P}_0$ (if $y=0$ was selected) or $\mathcal{P}_1$ (if $y=1$ was selected). We denote the mean of the $\mathcal{P}_0$ (resp. $\mathcal{P}_1$) distribution by $\boldsymbol{\mu}_0$ (resp. $\boldsymbol{\mu}_1$). The homogeneity of the linear classifier is without loss of much generality because we can assume $\boldsymbol{\mu}_0 = -\boldsymbol{\mu}_1$. We enforce this assumption, with minimal loss in accuracy, by recentering the data using a preliminary round of sampling (see Sec. 5).

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

$$\min_{\boldsymbol{\theta}\in\boldsymbol{R}^d} f(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{P}_*}[\ell(\boldsymbol{\xi}^T\boldsymbol{\theta}, 1)] = \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{P}_*}[-\boldsymbol{\xi}^T\boldsymbol{\theta} + \log(1+\exp(\boldsymbol{\xi}^T\boldsymbol{\theta}))]. \tag{2}$$

The most widely used method to solve (2) is SGD. Unlike gradient descent which uses the entire data to compute the gradient of the objective function, the SGD algorithm, at each iteration, generates a sample from the probability distribution and updates the iterate based only on this sample,

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \alpha\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}, 1), \tag{3}$$

where $\boldsymbol{\xi}_k \sim \mathcal{P}_*$. Our presentation of SGD assumes a constant step-size $\alpha > 0$. Constant step-size is commonly used in machine learning implementations despite the decreasing step-size often assumed to prove convergence (see, *e.g.*, [16]). With constant step-size, SGD is known to asymptotically converge to a neighborhood of the minimizer (see, *e.g.*, [15]). Yet, for binary classification, one does not require convergence to a minimizer in order to obtain good classifers.

It is known (see, *e.g.*, [14]) that 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 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.*, [8]). 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.

3.1. Stopping criterion

A common stopping criterion from deterministic first-order optimization methods is to terminate at an iterate satisfying $\|\nabla f(\boldsymbol{\theta})\|^2 < \varepsilon$ for a predetermined $\varepsilon > 0$. Yet, in stochastic optimization, the full gradient is inaccessible or it is simply too expensive to compute. Several works [1, 2, 5, 17, 18] 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. However, for binary classification, the minimizer of the loss function and a perfect classifier may not be the same or one may find a suitable substitute, at a lower cost, without having to compute the exact minimizer.

Optimal classifiers. In classification, we call a classifier, $\boldsymbol{\theta}^*$, *optimal* if it has the property that

$$\boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{P}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} > 0 \,|\, \boldsymbol{\xi} \sim \mathcal{P}_*\right), \tag{4}$$

i.e., the classifier, $\boldsymbol{\theta}^*$, minimizes the probability of misclassifying. Note there exist many optimal classifiers, in fact, the condition (4) is scale-invariant; hence, for any $\lambda > 0$, $\lambda \cdot \boldsymbol{\xi}^T\boldsymbol{\theta}^* > 0 \iff \boldsymbol{\xi}^T\boldsymbol{\theta}^* > 0$. Even though the binary classifier is scale-free, the logistic 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.

Termination test. Motivated by the above property of optimal classifiers, we propose the following termination test: Sample $\hat{\boldsymbol{\xi}}_k \sim \mathcal{P}_*$ and

$$\text{Terminate when } \ \hat{\boldsymbol{\xi}}_k^T\boldsymbol{\theta}_k \geq 1. \tag{5}$$

Algorithm 1: 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 $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\xi}_{\boldsymbol{k+1}}^T\boldsymbol{\theta}_k, 1)$ as in (3)
    Update $\boldsymbol{\theta}$ by setting
        $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \alpha\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\xi}_{\boldsymbol{k+1}}^T\boldsymbol{\theta}_k, 1)$    (6)
    Sample $\hat{\boldsymbol{\xi}}_{k+1} \sim \mathcal{P}_*$
    $k \leftarrow k + 1$
end
```

A second motivation for this termination test comes from support vector machine (SVM) theory [19] in which the scaling of the optimizing classifier is constrained so that the margin between classes is $O(1)$. Therefore, our termination test blends an SVM notion with SGD. Algorithm 1 describes the termination criterion (5) as applied with the update rule governed by SGD.

The termination test (5) 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. 5), we use the following termination test.

$$\text{Terminate when } \ \boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k \geq 1, \tag{7}$$

which imposes no computational overhead as SGD already computes $\boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k$. Unfortunately, we could not perform a straightforward analysis of (7) because it introduces additional dependencies in the sequences $\{\boldsymbol{\xi}_k\}_{k=1}^\infty$ and $\{\boldsymbol{\theta}_k\}_{k=0}^\infty$.

After testing both (5) and (7), we found that up to the noise from the randomness, their behaviors in numerical experiments were identical. The choice of 1 in (5) and (7) is arbitrary provided it is strictly positive. However, the constant should be positive enough to ensure a reasonable number of SGD iterations occur before termination. Our numerical experimentation (see Sec. 5) indicate the constant 1 worked well on a variety of real and simulated data.

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. 4, $\mathcal{P}_0 = N(\boldsymbol{\mu}_0, \sigma^2 I_d)$, $\mathcal{P}_1 = N(\boldsymbol{\mu}_1, \sigma^2 I_d)$, 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 $\sigma > 0$.

The minimizer of logistic regression. The analysis of our proposed stopping criteria in Algorithm 1 (see Sec. 4) involves knowing at least one optimal classifier. The following lemma (see Appendix. Sec. 3.1 for proof) provides an exact formula for an optimal classifier.

Lemma 1 *The function f defined in* (2) *has a unique minimizer at $\boldsymbol{\theta}^* = \frac{2\boldsymbol{\mu}}{\sigma^2}$. In addition, the set of optimal classifiers, in the sense of* (4)*, equals to $c \cdot \boldsymbol{\theta}^*$ for all $c > 0$.*

Therefore, up to rescaling the set of optimal classifiers is uniquely generated by $\frac{2\boldsymbol{\mu}}{\sigma^2}$. For the rest of the paper, we denote the vector $\boldsymbol{\theta}^*$ as the minimizer of the logistic loss.

4. Analysis of stopping criterion

In this section, we present our result on the stopping criterion (5) proposed in Sec. 3. Here we introduce the first iteration at which the stopping criterion is satisfied, denoted by the stopping time

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

The following theorem presents the bound on the expected number of iterations until our stopping criterion is satisfied. Proof is deferred to Appendix Sec. 4.1 and 4.2.

Theorem 2 *Let $\boldsymbol{\theta}_0 = \mathbf{0}$. Then the following are true*

1. *In the low variance regime* (i.e., $\sigma \leq 0.33\|\boldsymbol{\mu}\|$)*, for $b = \alpha\|\boldsymbol{\mu}\|^2$ and $M = 501 + 640\alpha\|\boldsymbol{\mu}\|^2$, it holds that*

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

2. *In the high variance regime* (i.e., $0.33\|\boldsymbol{\mu}\| \leq \sigma$)*, it holds that $\mathbb{E}[T] < +\infty$ provided the step-size α satisfies $\alpha \leq A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2(\|\boldsymbol{\mu}\|^2 + d\sigma^2)}$ for some positive constant A.*

In particular, 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.*, $\sigma \to 0$). It remains to determine whether the classifier at termination, $\boldsymbol{\theta}_T$, has a 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 $\mathbb{E}[|\boldsymbol{v}^T\boldsymbol{\theta}_T|]$ (see Appendix Sec. 4.3 for the proof).

Theorem 3 *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 estmiate holds:* $\mathbb{E}[|\boldsymbol{v}^T\boldsymbol{\theta}_T|] \leq \sigma\alpha\sqrt{\frac{2}{\pi}}\mathbb{E}[T]$.

Combining Theorem 2 and Theorem 3, for a fixed step-size α, we obtain an upper bound for $\mathbb{E}[|\boldsymbol{v}^T\boldsymbol{\theta}_T|]$. Therefore, in the low variance regime, as $\sigma \to 0$, $\mathbb{E}[|\boldsymbol{v}^T\boldsymbol{\theta}_T|]$ decreases exponentially whereas in the high variance regime, Theorem 3 yields a very loose bound. Yet despite this, our numerical results in Sec. 5 show promising accuracy of (5) in this case as well. We conjecture that the inequality can be significantly strengthened.

5. Numerical Experiments

We investigate the performance of our termination test on two popular data sets, MNIST [11] and CIFAR-10 [9], 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 (7); experiments using our test which required an extra sample (5) are not presented since the behaviors of the two criteria were indistinguishable on all data sets.

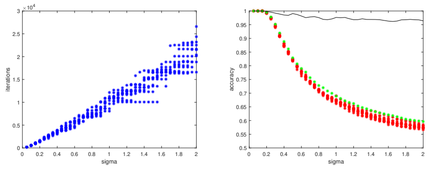


Figure 1. Performance of stopping criterion (7) on a mixture of Gaussians as σ is varied. Both plots show tests for values of σ equally spaced from 0.05 to 2.0. For each value of σ, 10 trials were run. The left plot shows the relationship between σ and k, the iteration number when (7) first holds. The right plot shows the accuracy as red asterisks. The green asterisks in the right plot 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 (7) holds to the accuracy of the optimal classifier.

Comparison with a popular stopping criterion. We include as a baseline a popular termination test, the small validation set (SVS) [10]. The SVS termination test is as follows. One fixes a validation set of p instances $(\zeta_1^{\mathrm{V}}, y_1^{\mathrm{V}}), \ldots, (\zeta_p^{\mathrm{V}}, y_p^{\mathrm{V}})$ drawn from the same distribution as the training data. Then on every l iterations, one checks the fraction correct of the current classifier $\boldsymbol{\theta}_{ml}$, where ml is the iteration index, on the p instances. If the fraction correct fails to increase compared to the last run of the SVS, then the iteration is 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 to be a multiple of p. In all of our tests of SVS, we chose $l = 2p$, which means that the additional cost of performing SVS on SGD is 50%. In contrast, the overhead for (7) is 0. The value of p is a tuning parameter for SVS; we exhibit results for three different choices of p.

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 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 $1/\sigma^2$. In all computational experiments, we choose $\alpha = 1/(16\tilde{\tau}^2)$ where $\tilde{\tau}^2$ 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 $\tilde{\tau}^2$ 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 the step-size is $\sigma^2 d$. The choice of 16 was manually tuned to get a good balance of performance versus accuracy on several test sets; refer to Appendix Sec. 7. A second hyperparameter also considered in that section is the '1' on the right-hand side of (7).

5.1. 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. 1, 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 (7) holds almost matches the optimal classifier; in particular, the averaged accuracy of our classifier/accuracy of the optimal classifier over the 10 runs, black curve in Fig. 1, never dips below 0.95.

In Fig. 2, the three plots on top, we compare performance of (7) 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 (7) holds (green '+') to test whether accuracy improves after termination.

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. 2, the three plots in bottom, shows our performance against SVS.

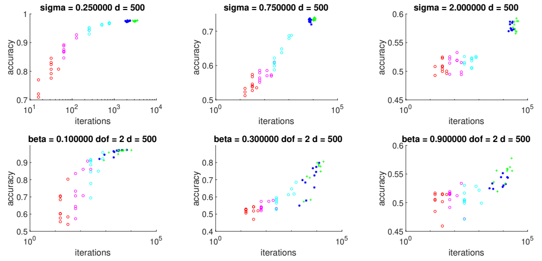


Figure 2. Each plot shows 10 random runs of SGD applied to the indicated data set, and for each of the ten runs, five termination tests corresponding to five colors were applied. SVS was tried with $p = 8, 32, 128$, depicted as red, magenta and cyan circles respectively. Test (7) is indicated with a blue asterisk. A green '+' corresponds to termination after $1.5k$ iterations, k is the iteration index (7) first holds.

5.2. Experiments with real data

MNIST handwritten digits. We compared our termination test on the MNIST handwritten digit set [11] ($d = 784$, no preprocessing of the data other than centering between the two means). Two trials are shown: distinguishing 1 from 8 (an easy case) and distinguishing 7 from 9 (a more difficult case). The test runs are obtained by running through the training data in different randomized orders.

CIFAR-10 image set. We compared our termination test on the CIFAR-10 [9] ($d = 3072$, no preprocessing of the data other than centering between the two means). 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.

Experimental conclusions. The plots in Figs. 2 and 3 show a consistent pattern that (7) achieves accuracy equal to or better than SVS. In the cases when the accuracy is equal, the iteration count for (7) is comparable or better. Iterating beyond the step for which (7) holds does not significantly improve accuracy. Another benefit of (7) apparent from all plots is that its behavior (in terms of number of iterations and accuracy) is more consistent across random trials, which is beneficial in the case that SGD is used as a subproblem of a larger computation.

Our computational experiments did not explore regularization via early stopping. Experiments showed that as SGD iterations continued, the accuracy on the test set eventually levels off but does not decrease significantly, *i.e.*, SGD for binary classification is not prone to overfitting. Because the test accuracy never shows marked decline, there is no opportunity for early stopping to regularize. However, we know of other settings in which early stopping has a strong regularizing effect (*e.g.*, conjugate gradient iterations for image deconvolution, already known in [20]), so if (7) is extended beyond binary classification in future work, there will likely also be an opportunity to explore regularization.

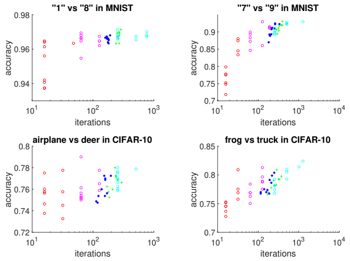


Figure 3. Results for MNIST and CIFAR-10 image classification data; refer to Fig. 2 for the legend.

References

[1] Nemirovski A., Juditsky A., Lan G., and Shapiro A. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

[2] Juditsky A. B., Nazin A. V., Nemirovsky A. S., and Tsybakov A. B. Algorithms of robust stochastic optimization based on mirror descent method. *preprint arXiv:1907.02707*, 2019.

[3] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[4] S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8 (3-4):231–357, 2015.

[5] Drusvyatskiy D. and Davis D. Robust stochastic optimization with the proximal point method. *preprint arXiv:1907.13307*, 2019.

[6] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, New York, NY, USA, 4th edition, 2010.

[7] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2016.

[8] Ziwei J. and Telgarsky M. Risk and parameter convergence of logistic regression. *preprint arXiv:1803.07300*, 2018.

[9] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[10] Prechelt L. *Early Stopping — But When?*, pages 53–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[11] 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.

[12] J. Lin and L. Rosasco. Optimal learning for multi-pass stochstic gradient methods. In *Advances in Neural Information Processing Systems (NeurIPs)*, pages 4556–4564, 2016.

[13] 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 (ICML)*, pages 2340–2348, 2016.

[14] Nacson M., Srebro N., and Soudry D. Stochastic Gradient Descent on Separable Data. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[15] G. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM J. Control Optim.*, 24(4):655–666, 1986.

[16] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[17] Ghadimi S. and Lan G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, i: a generic algorithmic framework. *SIAM J. Optim.*, 22 (4):1469–1492, 2012.

[18] Ghadimi S. and Lan G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: Shrinking procedures and optimal algorithms. *SIAM J. Optim.*, 23(4):2061–2089, 2013.

[19] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[20] A. van der Sluis and H. van der Vorst. SIRT-and CG-type methods for the iterative solution of sparse linear least-squares problems. *Linear Algebra Appl.*, 130:257–303, 1990.

[21] Yao Y., Rosasco L., and Caponnetto A. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

A termination criterion for SGD for binary classification Supplementary material

Sina Baghal SREZAZAD@UWATERLOO.CA
Courtney Paquette YUMIKO88@UW.EDU
Stephen Vavasis VAVASIS@UWATERLOO.CA
University of Waterloo, Canada

1. Introduction

Minimization of an expected loss objective function using linear predictors,

$$\min_{\boldsymbol{\theta}\in\boldsymbol{R}^d} f(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}}\ell(\boldsymbol{\zeta}^T\boldsymbol{\theta}, y), \tag{1}$$

is a central task in machine learning. Here the loss function $\ell : \boldsymbol{R}\times\boldsymbol{R}\to\boldsymbol{R}$, the probability distribution $\mathcal{P}$ is unknown, and the data sample $(\boldsymbol{\zeta}, y)\in\boldsymbol{R}^d\times\boldsymbol{R}$ is a random vector distributed as $\mathcal{P}$. The most prevalent algorithm employed for solving (1) is *stochastic gradient descent* (SGD). Whereas a significant amount of work has been devoted to the convergence analysis of SGD (see, *e.g.*, [4, 5, 23, 25]), leading, in particular, to learning rate schedules, the question of how to terminate the algorithm when one is near an optimal classifier remains largely unaddressed.

Yet, inexpensive stopping criteria are of utmost interest in machine learning. For instance, if one could produce a low cost test to determine near-optimality, then without sacrificing the quality of the solution or efficiency of the SGD algorithm, needless computational time would be eliminated. Secondly, early termination tests impose a degree of predictability on accuracy and running times– a useful quality when SGD occurs as a subproblem of a larger computation. Several works show that early termination of SGD can prevent overfitting, speed up learning procedures, and/or improve generalization properties [10, 14, 31]. Motivated by these facts, we sought to address from stochastic optimization the following question:

How to design a test to terminate SGD with a fixed learning rate that is inexpensive without sacrificing quality of the solution?

To do so, we simplified our setting to binary classification, one of the fundamental examples of supervised machine learning [28]. In binary classification, the learning algorithm is given a sequence of training examples $(\boldsymbol{\zeta}_1, y_1), (\boldsymbol{\zeta}_2, y_2), \ldots$, often noisy, where $\boldsymbol{\zeta}_i\in\boldsymbol{R}^d$ and $y_i\in\{0,1\}$ for each i. The job of the algorithm is to develop a rule for distinguishing future, unseen $\boldsymbol{\zeta}$'s that are classified as 1 from those classified as 0. In this work, we limit attention to linear classifiers. This means that the learning algorithm must determine a vector $\boldsymbol{\theta}$ such that the classification of $\boldsymbol{\zeta}$ is 1 when $\boldsymbol{\zeta}^T\boldsymbol{\theta} > 0$ else it is 0. Note that any algorithm for linear classification can be extended to one for nonlinear classification via the construction of "kernels"; see, *e.g.*, [28]. This extension is not pursued; we leave it for later work.

The usual technique for determining $\boldsymbol{\theta}$, which is also adopted herein, is to define a loss function that turns the discrete problem of computing a 1 or 0 for $\boldsymbol{\zeta}$ to a continuous quantity. Common choices of loss functions include logistic and hinge. For simplicity, we consider only the unregularized logistic and hinge loss in this work.

Our theoretical results assume that our data comes from a Gaussian mixture model (GMM). The GMM is attributed to [24]. The problem of identifying GMM parameters given random samples has attracted considerable attention in the literature; see, *e.g.*, the recent work of [2] and earlier references therein. Another common use of GMMs in the literature, similar to our application here, is as test-cases for a learning algorithm intended to solve a more general problem. Examples include clustering; see, *e.g.,* [12] and [22] and tensor factorization; see, *e.g.,* [29].

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.

Our contributions. In this paper, we introduce a new and simple termination criterion for stochastic gradient descent (SGD) applied to binary classification using logistic regression and hinge loss with constant step-size $\alpha > 0$. 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 exponentially 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.

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 [15] and CIFAR-10 [13]). In our experiments, we observe that our test yields relatively accurate classifiers with small variation across multiple runs.

Related works. To the best of our knowledge, the earliest comprehensive numerical testing of a stopping termination test for SGD in neural networks was introduced by [14]. His stopping criteria, which we denote as *small validation set* (SVS), periodically checks the iterate on a validation set.

Theoretical guarantees for SVS were established in the works of [17, 31]. [10] shows that SGD is uniformly stable and thus solutions with low training error found quickly generalize well. These results support exploring new computationally inexpensive termination tests– the spirit of this paper.

In a related topic, the relationship between generalization and optimization is an active area of research in machine learning. Much of the pioneering work in this area focused on understanding how early termination of algorithms, such as conjugate gradient, gradient descent, and SGD, can act as an implicit regularizer and thus exhibit better generalization properties [14, 16, 17, 30, 31]. The use of early stopping as a tool for improving generalization is not studied herein because our experiments indicate that for the problem under consideration, binary classification with a linear separator, the accuracy increases as SGD proceeds and ultimately reaches a steady value but does not decrease, meaning that there is no opportunity to improve generalization by stopping early. See also [1].

Instead of using a validation set to stop early, [8] employs an estimate of the marginal likelihood as a stopping criteria. Another termination test based upon a Wald-type statistic developed for solving least squares with reproducing kernels guarantees a minimax optimal testing [18]. However it is unclear the practical benefits of such procedures over a validation set.

Several works have introduced validation procedures to check the accuracy of solutions generated from stochastic algorithms based upon finding a point $\boldsymbol{\theta}_\varepsilon$ that satisfies a high confidence bound $\mathbb{P}(f(\boldsymbol{\theta}_\varepsilon) - \min f \le \varepsilon) \ge 1 - p$, in essence, using this as a stopping criteria (*e.g.*, see [1, 3, 6, 26, 27]). Yet, notably, all these procedures produce points with small function values. For binary classification, however, this could be quite expensive and a good classifier need not necessarily be the minimizer of the loss function. Ideally, one should terminate when the classifier's direction aligns with the optimal direction– the approach we pursue herein.

2. Background and preliminaries

Throughout we consider a Euclidean space, denoted by $\boldsymbol{R}^d$, with an inner product and an induced norm $\|\cdot\|$. The set of non-negative real numbers is denoted by $\boldsymbol{R}_{\ge 0}$. Bold-faced variables are vectors. Throughout, the matrix I_d is the d by d identity matrix. All stochastic quantities defined hereafter live on a probability space denoted by $(\mathbb{P}, \Omega, \mathcal{F})$, with probability measure $\mathbb{P}$ and the σ-algebra $\mathcal{F}$ containing subsets of Ω. Recall, a random variable (vector) is a measurable map from Ω to $\boldsymbol{R}$ ($\boldsymbol{R}^d$), respectively. An important example of a random variable is the *indicator of the event* $A \in \mathcal{F}$:

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

If X is a measurable function and $t \in \boldsymbol{R}$, we often simplify the notation for the pull back of the function X, to simply $\{\omega \in \Omega \;:\; X(\omega) \le t\} =: \{X \le t\}$. As is often in probability theory, we will not explicitly define the space Ω, but implicitly define it through random variables. For any sequence of random vectors $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_k)$, we denote the *$\sigma$-algebra generated by random vectors* $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_k$ by the notation $\sigma(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \dots, \boldsymbol{X}_k)$ and the *expected value of* $\boldsymbol{X}$ by $\mathbb{E}[\boldsymbol{X}] := \int_\Omega \boldsymbol{X}\, d\mathbb{P}$.

Particularly, we are interested in random variables that are distributed from normal distributions. In the next section, we state some known results about normal distributions.

Normal distributions The *probability density function of a univariate Gaussian* with mean μ and variance σ^2 is described by:

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

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

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

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 the paper. We list those which we specifically rely on. See [9] for proofs. Below, $\boldsymbol{v}, \boldsymbol{v}' \in \boldsymbol{R}^d$, $r \in \boldsymbol{R}$, $\boldsymbol{\xi} \sim N(\boldsymbol{\mu}, \sigma^2 I_d)$ and $\xi \sim N(\mu, \sigma^2)$. Also, $\psi \sim N(0,1)$.

Throughout our analysis, we encounter random variables of the form $\boldsymbol{v}^T\boldsymbol{\xi} + r$, i.e. affine transformations 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 other words, it holds that

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

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 below: The following is true

$$\boldsymbol{v}^T\boldsymbol{\xi} \text{ and } \boldsymbol{v}'^T\boldsymbol{\xi} \text{ are independent} \quad \text{ if and only if } \quad \boldsymbol{v}^T\boldsymbol{v}' = 0. \tag{3}$$

We will also use the following simple fact about truncated normal distributions:

$$\mathbb{E}_\xi[\xi 1_{\{\xi \leq b\}}] = 0 \Longrightarrow \Phi\left(\frac{b-\mu}{\sigma}\right) \cdot \exp\left(\frac{1}{2} \cdot \left(\frac{b-\mu}{\sigma}\right)^2\right) = \frac{\sigma}{\mu}. \tag{4}$$

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

$$\mathbb{E}\left[\|\boldsymbol{\xi}\|^2\right] = \|\boldsymbol{\mu}\|^2 + d\sigma^2, \quad \mathbb{E}_\xi[|\xi|] \leq \sqrt{\frac{2}{\pi}} \cdot \sigma + |\mu| \quad \text{and} \quad \mathbb{E}\left[|\psi|\right] = \sqrt{\frac{2}{\pi}}. \tag{5}$$

Martingales and stopping times Here we state some relevant definitions and theorems used in analyzing our stopping criteria in Section 4. We refer the reader to [7] for further details. 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 $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ 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

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

we say $\{X_k\}_{k=0}^{\infty}$ is a *supermartingale*. 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. 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 1 (See [7] 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 $\mathbb{E}\left[|X_{k+1} - X_k||\mathcal{F}_k\right] \leq B$ a.s. for some constant $B > 0$, then it holds that $\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$.*

As we illustrate in Section 4, a connection between stopping criteria (i.e. the decision to stop an algorithm) and stopping times naturally exists.

3. Stopping criterion for stochastic gradient descent

We analyze learning by minimizing an expected loss problem of homogeneous 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. Here the samples $(\boldsymbol{\zeta}, y) \in \boldsymbol{R}^d \times \{0, 1\}$. We recall that in logistic regression the loss function is defined as follows

$$\ell(x, y) := -yx + \log\left(1 + \exp(x)\right). \tag{6}$$

Also, the hinge loss is defined as the following

$$\ell(x, y) := \begin{cases} \max(1 - x, 0) & y = 1, \\ \max(1 + x, 0) & y = 0. \end{cases} \tag{7}$$

The data comes from a mixture model, that is, flip a coin to determine whether an item is in the $y = 0$ or $y = 1$ class, then generate the sample $\boldsymbol{\zeta}$ from either the distribution $\mathcal{P}_0$ (if $y = 0$ was selected) or $\mathcal{P}_1$ (if $y = 1$ was selected). We denote the mean of the $\mathcal{P}_0$ (resp. $\mathcal{P}_1$) distribution by $\boldsymbol{\mu}_0$ (resp. $\boldsymbol{\mu}_1$). The homogeneity of the linear classifier is without loss of much generality because we can assume $\boldsymbol{\mu}_0 = -\boldsymbol{\mu}_1$. We enforce this assumption, with minimal loss in accuracy, by recentering the data using a preliminary round of sampling (see Sec. 5).

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

$$\min_{\boldsymbol{\theta} \in \boldsymbol{R}^d} f(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{P}_*}[\ell(\boldsymbol{\xi}^T \boldsymbol{\theta}, 1)] \tag{8}$$

Let us remark that the right-hand side of (8) 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 (8) to be over $\mathbb{R}^d \setminus \left\{\boldsymbol{\xi}_k : \boldsymbol{\xi}_k^T \boldsymbol{\theta}_{k-1} = 1\right\}$ on which the argument is differentiable with respect to $\boldsymbol{\theta}_{k-1}$.

The most widely used method to solve (8) is SGD. Unlike gradient descent which uses the entire data to compute the gradient of the objective function, the SGD algorithm, at each iteration, generates a sample from the probability distribution and updates the iterate based only on this sample,

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \alpha \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\xi}_k^T \boldsymbol{\theta}_{k-1}, 1), \tag{9}$$

where $\boldsymbol{\xi}_k \sim \mathcal{P}_*$. Our presentation of SGD assumes a constant step-size $\alpha > 0$. Constant step-size is commonly used in machine learning implementations despite the decreasing step-size often assumed to prove convergence (see, *e.g.*, [25]). [1] explain in more detail the theoretical basis for both constant and decreasing step-size and provide an explanation as well as workarounds for the poor practical performance of decreasing step-size. However, in practice, constant step-size is still widely used. With constant step-size, SGD is known to asymptotically converge to a neighborhood of the minimizer (see, *e.g.*, [23]). Yet, for binary classification, one does not require convergence to a minimizer in order to obtain good classifiers.

For homogeneous linear classifiers applied to the hinge loss function, it has been shown ([21]) that the homotopic sub-gradient method converges to a maximal margin solution on linearly separable data. In ([19]), 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.*, [11]). 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.

3.1. Stopping criterion

A common stopping criterion from deterministic first-order optimization methods is to terminate at an iterate satisfying $\|\nabla f(\boldsymbol{\theta})\|^2 < \varepsilon$ for a predetermined $\varepsilon > 0$. Yet, in stochastic optimization, the full gradient is inaccessible or it is simply too expensive to compute. Several works [1, 3, 6, 26, 27] 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. However, for binary classification, the minimizer of the loss function and a perfect classifier may not be the same or one may find a suitable substitute, at a lower cost, without having to compute the exact minimizer.

Optimal classifiers. In classification, we call a classifier, $\boldsymbol{\theta}^*$, *optimal* if it has the property that

$$\boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathbb{P}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} > 0 \,|\, \boldsymbol{\xi} \sim \mathcal{P}_*\right), \tag{10}$$

i.e., the classifier, $\boldsymbol{\theta}^*$, minimizes the probability of misclassifying. Note there exist many optimal classifiers, in fact, the condition (10) is scale-invariant; hence, for any $\lambda > 0$, $\lambda \cdot \boldsymbol{\xi}^T\boldsymbol{\theta}^* > 0 \iff \boldsymbol{\xi}^T\boldsymbol{\theta}^* > 0$. 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.

Termination test. Motivated by the above property of optimal classifiers, we propose the following termination test: Sample $\hat{\boldsymbol{\xi}}_k \sim \mathcal{P}_*$ and

$$\text{Terminate when } \; \hat{\boldsymbol{\xi}}_k^T\boldsymbol{\theta}_k \geq 1. \tag{11}$$

A second motivation for this termination test comes from support vector machine (SVM) theory [28] in which the scaling of the optimizing classifier is constrained so that the margin between classes is $O(1)$. Therefore, our termination test blends an SVM notion with SGD. Algorithm 1 describes the termination criteria (11) as applied with the update rule governed by SGD.

The termination test (11) 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. 5), we use the following termination test.

$$\text{Terminate when } \; \boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k \geq 1, \tag{12}$$

which imposes no computational overhead as SGD already computes $\boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k$. Unfortunately, we could not perform a straightforward analysis of (12) because it introduces additional dependencies in the sequences $\{\boldsymbol{\xi}_k\}_{k=1}^\infty$ and $\{\boldsymbol{\theta}_k\}_{k=0}^\infty$. After testing both (11) and (12), we found that up to the noise from the randomness, their behaviors in numerical experiments were identical.

```
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 $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\xi}_{\boldsymbol{k+1}}^T\boldsymbol{\theta}_k, 1)$ as in (9)
    Update $\boldsymbol{\theta}$ by setting
        $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \alpha\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\xi}_{\boldsymbol{k+1}}^T\boldsymbol{\theta}_k, 1)$    (13)
    Sample $\hat{\boldsymbol{\xi}}_{k+1} \sim \mathcal{P}_*$
    $k \leftarrow k + 1$
end
```

Algorithm 1: SGD with termination test

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. 4, $\mathcal{P}_0 = N(\boldsymbol{\mu}_0, \sigma^2 I_d)$, $\mathcal{P}_1 = N(\boldsymbol{\mu}_1, \sigma^2 I_d)$, 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 $\sigma > 0$.

The minimizer of logistic and hinge regression In (10) we defined $\boldsymbol{\theta}^*$ to be any member of the set of optimal classifiers. For the remainder of this section, we provide an exact characterization of this set. In the next lemma, we redefine $\boldsymbol{\theta}^*$ to the minimizer of the expected loss function for either hinge or logistic and show that it is a positive scalar multiple of $\boldsymbol{\mu}$. We will continue to use $\boldsymbol{\theta}^*$ with this meaning for the remainder of the paper. In the lemma after that, we show that the set of optimal classifiers are exactly positive scalar multiples of $\boldsymbol{\mu}$ (or of $\boldsymbol{\theta}^*$).

Lemma 2 (Minimizer of the logistic and hinge loss) *The function f defined in (8) with ℓ defined in (6) or (7) 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*

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{\sigma}{\|\boldsymbol{\mu}\|} = \Phi(w) \cdot \exp(\tfrac{1}{2}w^2). \tag{14}$$

Proof

We consider the logistic and hinge loss case separately.

1. **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

$$\boldsymbol{v}^T\nabla f\left(\boldsymbol{\theta}\right) = \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\boldsymbol{\xi}^T\boldsymbol{v}}{1+\exp(\boldsymbol{\xi}^T\boldsymbol{\theta})}\right] = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\xi}^T\boldsymbol{v}]\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{1+\exp(\boldsymbol{\xi}^T\boldsymbol{\theta})}\right] = \boldsymbol{v}^T\boldsymbol{\mu}\cdot\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{1+\exp(\boldsymbol{\xi}^T\boldsymbol{\theta})}\right]. \tag{15}$$

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 (15), 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 (15) 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 (15), 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), $z := \boldsymbol{\mu}^T\boldsymbol{\xi} \sim N(\|\boldsymbol{\mu}\|^2, \sigma^2\|\boldsymbol{\mu}\|^2)$. 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

$$\frac{z}{1+\exp(\rho z)}\exp\left(-\frac{(z-\|\boldsymbol{\mu}\|^2)^2}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right)=\frac{z}{2\cosh\left(\frac{\rho z}{2}\right)}\exp\left(-\frac{\rho z}{2}-\frac{(z-\|\boldsymbol{\mu}\|^2)^2}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right)$$
$$=\frac{z}{2\cosh\left(\frac{\rho z}{2}\right)}\exp\left(-\frac{z^2+\left(\rho\sigma^2\|\boldsymbol{\mu}\|^2-2\|\boldsymbol{\mu}\|^2\right)z+\|\boldsymbol{\mu}\|^4}{2\sigma^2\|\boldsymbol{\mu}\|^2}\right). \tag{16}$$

When $\rho=\rho^*$, we observe that equation (16) is an odd function of z. Therefore, the function $g(\rho^*)=0$, i.e. the integral of (16) 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.

2. **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

$$-\boldsymbol{v}^T\nabla f(\boldsymbol{\theta})=\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{v}^T\boldsymbol{\xi}1_{\{\boldsymbol{\xi}^T\boldsymbol{\theta}\le 1\}}]=\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{v}^T\boldsymbol{\xi}]\mathbb{E}_{\boldsymbol{\xi}}[1_{\{\boldsymbol{\xi}^T\boldsymbol{\theta}\le 1\}}]=\boldsymbol{v}^T\boldsymbol{\mu}\cdot\mathbb{E}_{\boldsymbol{\xi}}[1_{\{\boldsymbol{\xi}^T\boldsymbol{\theta}\le 1\}}]. \tag{17}$$

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 (17), 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 (17) 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 (17), we deduce $\|\boldsymbol{v}\|^2=\boldsymbol{v}^T\nabla f(\rho\boldsymbol{\mu})=0$ yielding a contradiction.

Next, consider the function

$$g(\rho)=\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\mu}^T\boldsymbol{\xi}1_{\{\rho\boldsymbol{\xi}^T\boldsymbol{\mu}\le 1\}}]. \tag{18}$$

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

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

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

$$\boldsymbol{\mu}^T\boldsymbol{\xi}\left(1_{\{\rho_1\boldsymbol{\mu}^T\boldsymbol{\xi}\le 1\}}-1_{\{\rho_2\boldsymbol{\mu}^T\boldsymbol{\xi}\le 1\}}\right)\ge 0\quad\text{for any value of }\boldsymbol{\xi}. \tag{19}$$

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 (19) holds with equality, almost surely. Clearly, this yields a contradiction. It remains to show (14). By (18), 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) and (4), we obtain that

$$\Phi\left(\frac{1-\rho^*\|\boldsymbol{\mu}\|^2}{\rho^*\sigma\|\boldsymbol{\mu}\|}\right)\cdot\exp\left(\frac{1}{2}\cdot\left(\frac{1-\rho^*\|\boldsymbol{\mu}\|^2}{\rho^*\sigma\|\boldsymbol{\mu}\|}\right)^2\right) = \frac{1}{\sqrt{2\pi}}\cdot\frac{\sigma}{\|\mu\|}. \tag{20}$$

The result immediately follows.

■

The previous lemma has defined $\boldsymbol{\theta}^*$ to be the minimizer of the loss function and showed that it is a positive multiple of $\boldsymbol{\mu}$. We now show that this $\boldsymbol{\theta}^*$ and its positive scalar multiples are exactly the set of optimal classifiers in the sense of (10), i.e., we give an exact characterization of that set.

Lemma 3 (Characterization of the optimal classifier) *The following is true*

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\,\mathbb{P}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} > 0\right) = \{\lambda\cdot\boldsymbol{\theta}^* : \lambda > 0\}. \tag{21}$$

Proof Observe that the following simple fact holds.

$$\mathbb{P}_{\hat{\boldsymbol{\xi}}}\left(\hat{\boldsymbol{\xi}}^T\boldsymbol{\theta} \geq t\right) = \Phi^c\left(\frac{\boldsymbol{\mu}^T\boldsymbol{\theta} - t}{\sigma\|\boldsymbol{\theta}\|}\right), \quad \text{for all } \boldsymbol{\theta}\in\boldsymbol{R}^d,\, t\in\boldsymbol{R} \text{ and } \hat{\boldsymbol{\xi}}\sim N(\boldsymbol{\mu},\sigma^2 I_d). \tag{22}$$

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. ■

4. Analysis of stopping criterion

In this section, we present our analysis of the stopping criterion (11) proposed in Section 3. Here we introduce the first iteration at which the stopping criterion is satisfied, denoted by the random variable

$$T := \inf\left\{k > 0 : \hat{\boldsymbol{\xi}}_k^T\boldsymbol{\theta}_k \geq 1\right\}. \tag{23}$$

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.

$$\mathcal{F}_0 = \sigma(\boldsymbol{\theta}_0) \quad \text{and} \quad \mathcal{F}_k := \sigma(\boldsymbol{\theta}_0, \hat{\boldsymbol{\xi}}_1, \boldsymbol{\xi}_1, \hat{\boldsymbol{\xi}}_2, \boldsymbol{\xi}_2, \ldots, \hat{\boldsymbol{\xi}}_k, \boldsymbol{\xi}_k), \quad \text{for all } k \geq 1 \tag{24}$$

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.* (11)) 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 $\delta > 0$,

$$\exists\, \delta > 0 \text{ such that } \mathbb{P}_{\hat{\boldsymbol{\xi}}}\left(\hat{\boldsymbol{\xi}}^T \boldsymbol{\theta} \geq 1\right) \geq \delta. \tag{25}$$

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

$$(\mathbb{E}[V(\boldsymbol{\theta}_k)|\mathcal{F}_{k-1}] - V(\boldsymbol{\theta}_{k-1})])1_{\{\boldsymbol{\theta}_{k-1} \notin C\}} \leq -b1_{\{\boldsymbol{\theta}_{k-1} \notin C\}} \tag{26}$$

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, see Lemma 9 below.

A natural choice for the target set is a neighborhood of the unique optimum solution of (8), $\boldsymbol{\theta}^*$, with the drift function $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$. Indeed, it is known the iterates of SGD converge to a neighborhood of $\boldsymbol{\theta}^*$ ([23]). 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

   $$C = \{\boldsymbol{\theta} : \boldsymbol{\mu}^T \boldsymbol{\theta} \geq 1\}, \tag{27}$$

   and the drift function by

   $$V(\boldsymbol{\theta}) = \left(M - \boldsymbol{\mu}^T \boldsymbol{\theta}\right)^2, \tag{28}$$

   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

   $$C = \{\boldsymbol{\theta} : |\rho\sigma^2 - 1| < 1 \text{ and } \sigma\|\tilde{\boldsymbol{\theta}}\| \leq c'\}, \tag{29}$$

   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

   $$V(\boldsymbol{\theta}) = \frac{1}{2\alpha}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2. \tag{30}$$

In Section 4.1 (resp. Section 4.2) we show that the pairs (C, V) defined in (27) and (28) (resp. (29) and (30)) satisfies the drift equation (26) 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:

$$\hat{\boldsymbol{\xi}}_k^T \boldsymbol{\theta}_k \geq 1 \text{ and } \boldsymbol{\theta}_k \in C. \tag{31}$$

The SGD algorithm does not know the value of $\boldsymbol{\theta}^*$; therefore at each iteration, it cannot check whether the condition (31) occurs. Nevertheless, we are able to compute a bound on the average waiting time until (31) holds and the first time (31) holds is always an upper bound on T, our stopping criterion. This is summarized in Lemma 4. Precisely, if we denote by

$$T_C := \inf\{k > 0 : \hat{\boldsymbol{\xi}}_k^T \boldsymbol{\theta}_k \geq 1 \text{ and } \boldsymbol{\theta}_k \in C\}, \tag{32}$$

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

$$\tau_1 := \inf\{k > 0 : \boldsymbol{\theta}_k \in C\} \tag{33}$$

and inductively, for $m \geq 2$,

$$\tau_m := \inf\{k > \tau_{m-1} : \boldsymbol{\theta}_k \in C\}. \tag{34}$$

The following lemma formalizes the discussion above.

Lemma 4 *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*

$$\mathbb{E}[T] \leq \mathbb{E}[T_C] \leq \sum_{m=1}^{\infty} \mathbb{E}[\tau_m](1-\delta)^{m-1}, \tag{35}$$

where δ satisfies (25).

Proof We first show that

$$\mathbb{E}\left[1_{\{T_C \geq \tau_m\}}\right] \leq (1-\delta)^{m-1}. \tag{36}$$

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 (36) holds. Now since $\mathbb{E}[1_{\{T_C=+\infty\}}] \le \mathbb{E}[1_{\{T_C\ge\tau_m\}}]$ for all $m \ge 1$, it follows from (36) that $T_C < \infty$ a.s. We next observe that

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

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

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

The proof is complete. ∎

Now, in view of Lemma 4, it suffices to bound $\mathbb{E}[\tau_m]$ by a sequence which can not grow too fast in m. Indeed, we show that (26) implies the following

$$\mathbb{E}[\tau_m] = \mathcal{O}(m). \tag{37}$$

Theorem 5 *(Low Regime) Let $\{\boldsymbol{\theta}_k\}_{k=0}^{\infty}$ be a sequence generated by Algorithm 1 such that $\boldsymbol{\theta}_0 = \mathbf{0}$. There exists positive constants c, b and M such that provided $\sigma \le c\|\boldsymbol{\mu}\|$ the following holds.*

$$\mathbb{E}[T] \le 2 + \frac{2M^2}{b} \cdot \left(\Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3}{\|\boldsymbol{\mu}\|} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right) + 1\right). \tag{38}$$

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

1. *For the logistic loss,*

$$c = 0.33, \quad b = \alpha\|\boldsymbol{\mu}\|^2, \quad \textit{and } M = 501 + 640\alpha\|\boldsymbol{\mu}\|^2. \tag{39}$$

2. *For the hinge loss,*

$$c = 1.25, \quad b = \alpha\|\boldsymbol{\mu}\|^2, \quad \textit{and } M = 501 + 782\alpha\|\boldsymbol{\mu}\|^2. \tag{40}$$

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 5 in Section 4.3. 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 6 *(High Regime) Suppose that $\sigma > c\|\boldsymbol{\mu}\|$ where c is defined in (39) and (40). Then there exists a universal positive constant A such that if the step-size α satisfies*

$$\alpha \leq A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2(\|\boldsymbol{\mu}\|^2 + d\sigma^2)}, \tag{41}$$

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 7 *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*

$$\mathbb{E}[|\boldsymbol{v}^T\boldsymbol{\theta}_T|] \leq \sigma\alpha\sqrt{\frac{2}{\pi}}\mathbb{E}[T]. \tag{42}$$

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 6 yields a very loose bound. Yet despite this, our numerical result in Section 5 show promising accuracy of (11) in this case as well. We conjecture that the inequality can be significantly strengthened.

4.1. Low regime, proof of Theorem 5

In this section, we investigate the low variance regime. We consider the target set C and function V defined in (27) and (28) respectively, *i.e.*

$$C = \{\boldsymbol{\theta} : \boldsymbol{\mu}^T\boldsymbol{\theta} \geq 1\}, \quad V(\boldsymbol{\theta}) = \left(M - \boldsymbol{\mu}^T\boldsymbol{\theta}\right)^2, \tag{43}$$

where M is a constant to be determined. Next lemma shows that the drift equation (26) holds for the pair (C, V).

Lemma 8 (Drift equation) *Consider the SGD algorithm and let the set C and the function V be as in (43). Define the constants c, b, M as in (39) and (40). 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 (26) with the constant b.*

Proof For simplicity we write $\mathcal{F}_{-1} := \sigma\left(\{\boldsymbol{\theta}_0 = \boldsymbol{\theta}\}\right)$. 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 (26), 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

$$V(\boldsymbol{\theta}_k) = V(\boldsymbol{\theta}_{k-1}) - \frac{2\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_k(M - \boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1})}{1 + \exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1})} + \frac{\alpha^2(\boldsymbol{\mu}^T\boldsymbol{\xi}_k)^2}{(1 + \exp(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}))^2}. \tag{44}$$

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 (5), and the second to last from the observation that for any X normally distributed, $\mathbb{P}(\mathbb{E}[X]\geq X)=1/2$ and $\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\sim N(\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1},\sigma^2\left\|\boldsymbol{\theta}_{k-1}\right\|^2)$ and $\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1}<1$. The last inequality uses the assumption $\sigma\leq 0.33\left\|\boldsymbol{\mu}\right\|$. By taking the conditional expectations of (44) combined with the above sequence of inequalities, we deduce the following bound

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

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 (5). A quick computation after plugging in the value of M and the bound $\sigma\leq 0.33\|\boldsymbol{\mu}\|$ from (39) yields the drift equation (26) with $b=\alpha\|\boldsymbol{\mu}\|^2$.

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

$$
V(\boldsymbol{\theta}_k)=V(\boldsymbol{\theta}_{k-1})-2\alpha(M-\boldsymbol{\mu}^T\boldsymbol{\theta}_{k-1})\boldsymbol{\mu}^T\boldsymbol{\xi}_k 1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}+\alpha^2(\boldsymbol{\mu}^T\boldsymbol{\xi}_k)^2 1_{\{\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1}\leq 1\}}. \tag{45}
$$

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 (5). The last inequality uses the assumption $\sigma \leq 1.25\|\boldsymbol{\mu}\|$. By taking conditional expectations of (45) combined with the above sequence of inequalities, we deduce the bound

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

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

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

Lemma 9 ([20], Theorem 11.3.4) *Suppose that $V : \boldsymbol{R}^d \to [0, +\infty)$ is a drift function with respect to some target set C i.e. for some constant $b \in (0, +\infty)$ the drift equation (26) holds. The following is true*

$$\mathbb{E}[\tilde{\tau}_1|\boldsymbol{\theta}_0 = \boldsymbol{\theta}] \leq \tfrac{1}{b}V(\boldsymbol{\theta}). \tag{46}$$

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

Proposition 10 *(Bound on $\mathbb{E}[\tau_m]$) Let $\boldsymbol{\theta}_0 = \boldsymbol{0}$ and assume the notation and assumptions of Lemma 8 hold. The following is true for all $m \geq 1$*

$$\mathbb{E}[\tau_m] \leq (m-1)\left(1 + \frac{M^2}{b}\cdot\Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3 M^2}{\|\boldsymbol{\mu}\|b}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right)\right) + \frac{M^2}{b}. \tag{47}$$

Proof First, the result for $m = 1$ follows immediately by combining Lemmas 8 and 9 with $\boldsymbol{\theta}_0 = \boldsymbol{0}$. 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

$$\begin{aligned}\mathbb{E}\big[(\tau_m - \tau_{m-1})\wedge n|\mathcal{F}_{\tau_{m-1}+1}\big] &= \mathbb{E}\left[((\tau_m - \tau_{m-1})\wedge n)|\mathcal{F}_{\tau_{m-1}+1}\right]1_{\{\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\geq 1\}} \\ &\qquad + \mathbb{E}\left[((\tau_m - \tau_{m-1})\wedge n)|\mathcal{F}_{\tau_{m-1}+1}\right]1_{\{\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}<1\}} \\ &= 1_{\{\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\geq 1\}} + \mathbb{E}\left[((\tau_m - \tau_{m-1})\wedge n)|\mathcal{F}_{\tau_{m-1}+1}\right]1_{\{\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}<1\}} \\ &= 1_{\{\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\geq 1\}} + \sum_{i=1}^{\infty}\mathbb{E}\left[(\tau_m - \tau_{m-1})\wedge n|\mathcal{F}_{\tau_{m-1}+1}\right]1_{\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\leq i\}} \\ &= 1 + \sum_{i=1}^{\infty}\mathbb{E}\left[\tilde{\tau}_1\wedge n|\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\tau_{m-1}+1}\right]1_{\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\leq i\}}.\end{aligned} \tag{48}$$

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

$$1_{\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\leq i\}} \leq 1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}<\frac{1-i}{\alpha}\}}. \tag{49}$$

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

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

where the second inequality follows because $\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}} \geq 1$ and the last inequality because

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

$$\begin{aligned}
1\left\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\leq i\right\} &\leq 1\left\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\right\} \\
&\leq 1\left\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}} - \alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}1_{\{\boldsymbol{\xi}^T_{\tau_{m-1}+1}\boldsymbol{\theta}_{\tau_{m-1}}\leq 1\}}\right\} \\
&\leq 1\left\{i-1<-\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}1_{\{\boldsymbol{\xi}^T_{\tau_{m-1}+1}\boldsymbol{\theta}_{\tau_{m-1}}\leq 1\}}\right\} \\
&= 1\left\{i-1<-\alpha\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}\right\}.
\end{aligned} \tag{50}$$

Therefore we have shown that (49) holds. Setting $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\tau_{m-1}+1}$, by Lemma 9 for each $i \geq 1$, we deduce

$$\begin{aligned}
\mathbb{E}\left[\tilde{\tau}_1 \wedge n | \boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\tau_{m-1}+1}\right] 1_{\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\leq i\}} &\leq \frac{(M-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1})^2}{b} 1_{\{i-1<1-\boldsymbol{\mu}^T\boldsymbol{\theta}_{\tau_{m-1}+1}\leq i\}} \\
&\leq \frac{(M+i-1)^2}{b} 1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}<\frac{1-i}{\alpha}\}}.
\end{aligned} \tag{51}$$

Finally we observe that

$$\begin{aligned}
\mathbb{E}\left[1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{\tau_{m-1}+1}<\frac{1-i}{\alpha}\}}\right] &= \mathbb{E}\left[\sum_{k=1}^{\infty} 1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{k+1}<\frac{1-i}{\alpha}\}} 1_{\{\tau_{m-1}=k\}}\right] \\
&= \sum_{k=1}^{\infty} \mathbb{E}\left[1_{\{\boldsymbol{\mu}^T\boldsymbol{\xi}_{k+1}<\frac{1-i}{\alpha}\}}\right] \mathbb{E}\left[1_{\{\tau_{m-1}=k\}}\right] \\
&= \Phi\left(\frac{\frac{1-i}{\alpha} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}\right) \sum_{k=1}^{\infty} \mathbb{E}\left[1_{\{\tau_{m-1}=k\}}\right] \\
&= \Phi\left(\frac{\frac{1-i}{\alpha} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}\right).
\end{aligned} \tag{52}$$

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 (48), (51), and (52), we obtain the following

$$\begin{aligned}\mathbb{E}\big[(\tau_m - \tau_{m-1}) \wedge n\big] &\le 1 + \frac{M^2}{b} \cdot \Phi\left(-\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \sum_{i=2}^{\infty} \frac{(M+i-1)^2}{b} \cdot \Phi\left(\frac{\frac{1-i}{\alpha} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}\right) \\ &= 1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \sum_{i=2}^{\infty} \frac{(M+i-1)^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|^2 + \frac{i-1}{\alpha}}{\sigma\|\boldsymbol{\mu}\|}\right) \\ &\le 1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma\|\boldsymbol{\mu}\|}{b\sqrt{2\pi}} \cdot \sum_{i=2}^{\infty} \frac{(M+i-1)^2}{\alpha\|\boldsymbol{\mu}\|^2 + i - 1} \cdot \exp\left(-\frac{1}{2}\left(\frac{\|\boldsymbol{\mu}\|^2 + \frac{i-1}{\alpha}}{\sigma\|\boldsymbol{\mu}\|}\right)^2\right),\end{aligned} \tag{53}$$

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 $\frac{M+i-1}{\alpha\|\boldsymbol{\mu}\|^2+i-1} \le \frac{M}{\alpha\|\boldsymbol{\mu}\|^2}$ holds for all $i \ge 2$. Using this we obtain the following bound

$$\begin{aligned}&\sum_{i=2}^{\infty} \frac{(M+i-1)^2}{\alpha\|\boldsymbol{\mu}\|^2 + i - 1} \cdot \exp\left(-\frac{1}{2}\left(\frac{\|\boldsymbol{\mu}\|^2 + \frac{i-1}{\alpha}}{\sigma\|\boldsymbol{\mu}\|}\right)^2\right) \\ &\qquad\le \frac{\sigma M^2}{\alpha\|\boldsymbol{\mu}\|^3} \cdot \sum_{i=2}^{\infty} \frac{\alpha\|\boldsymbol{\mu}\|^2 + i - 1}{\alpha\sigma\|\boldsymbol{\mu}\|} \cdot \exp\left(-\frac{1}{2}\left(\frac{\alpha\|\boldsymbol{\mu}\|^2 + i - 1}{\alpha\sigma\|\boldsymbol{\mu}\|}\right)^2\right) \\ &\qquad\le \frac{\sigma M^2}{\alpha\|\boldsymbol{\mu}\|^3} \cdot \alpha\sigma\|\boldsymbol{\mu}\| \cdot \int_{\frac{\|\boldsymbol{\mu}\|}{\sigma}}^{+\infty} t\exp\left(-\frac{t^2}{2}\right) dt \\ &\qquad= \frac{\sigma^2 M^2}{\|\boldsymbol{\mu}\|^2} \cdot \exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right).\end{aligned} \tag{54}$$

Here we have used that $t \mapsto t\exp(-\frac{t^2}{2})$ is decreasing over $[1, +\infty)$. Combining (53) and (54), we obtain that

$$\mathbb{E}\left[(\tau_m - \tau_{m-1}) \wedge n\right] \le 1 + \frac{M^2}{b} \cdot \Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3 M^2}{\|\boldsymbol{\mu}\|b} \cdot \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right). \tag{55}$$

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

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

We then iterate the above inequality yielding

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

The result follows by plugging in the bound from Lemma 9 for the base case $m = 1$. ■

We are now ready to prove Theorem 5.

Proof [Proof of Theorem 5] 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 $\mathbb{P}_{\hat{\boldsymbol{\xi}}\sim N(\boldsymbol{\mu},\sigma^2 I_d)}\left(\hat{\boldsymbol{\xi}}^T\boldsymbol{\theta}\geq 1\right)\geq \frac{1}{2}$ for any $\boldsymbol{\theta}\in C$. Therefore $\delta=\frac{1}{2}$ satisfies (25). By Proposition 10 with Lemma 4, we conclude that

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

■

4.2. High regime, proof of Theorem 6

In this section, we consider the high variance regime. We consider the target set C and the function V defined in (29) and (30), respectively, *i.e.*

$$C:=\left\{\boldsymbol{\theta}:|\rho-\rho^*|<\tfrac{1}{2}\rho^* \text{ and } \sigma\|\tilde{\boldsymbol{\theta}}\|\leq c'\right\} \quad \text{and} \quad V(\boldsymbol{\theta}):=\frac{1}{2\alpha}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2, \tag{56}$$

where the minimizer $\boldsymbol{\theta}^*=\rho^*\boldsymbol{\mu}$ is defined in Lemma 2 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.

Lemma 11 *Consider the optimization problem* (8) *where* $\ell:\boldsymbol{R}\times\boldsymbol{R}\to\boldsymbol{R}$ *is either the logistic or hinge loss function. Denote the vector* $\boldsymbol{\theta}^*$ *as the unique minimizer of* f *in* (8)*. Let* $\boldsymbol{\theta}_0\in\boldsymbol{R}^d$*. The sequence* $\{\boldsymbol{\theta}_k\}_{k=0}^{\infty}$ *generated by SGD satisfies the following for all* $k\geq 1$*,*

$$f(\boldsymbol{\theta}_{k-1})-f(\boldsymbol{\theta}^*)\leq \frac{1}{2\alpha}\left(\|\boldsymbol{\theta}_{k-1}-\boldsymbol{\theta}^*\|^2-\mathbb{E}\left[\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^*\|^2|\mathcal{F}_{k-1}\right]\right)+\frac{\alpha}{2}\left(\|\boldsymbol{\mu}\|^2+d\sigma^2\right). \tag{57}$$

Proof Define the quantity

$$\boldsymbol{g_k}:=\frac{1}{\alpha}\left(\boldsymbol{\theta}_{k-1}-\boldsymbol{\theta}_k\right)=\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1},1\right).$$

Here, it is easy to check that the derivative with respect to $\boldsymbol{\theta}$ and the expectation over $\boldsymbol{\xi}_k$ are interchangeable, thus yielding

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

By convexity of the function f, we have the following

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

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

$$f(\boldsymbol{\theta}_{k-1})-f(\boldsymbol{\theta}^*)\leq\frac{1}{2\alpha}\left(\|\boldsymbol{\theta}_{k-1}-\boldsymbol{\theta}^*\|^2-\mathbb{E}\left[\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^*\|^2|\mathcal{F}_{k-1}\right]\right)+\frac{\alpha}{2}\mathbb{E}_{\boldsymbol{\xi}_k}\left[\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1},1)\|^2\right]. \tag{58}$$

We next observe the following bound

$$\mathbb{E}_{\boldsymbol{\xi}_k}[\|\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\xi}_k^T\boldsymbol{\theta}_{k-1},1\right)\|^2\,|\,\mathcal{F}_{k-1}] \le \mathbb{E}_{\boldsymbol{\xi}_k}[\|\boldsymbol{\xi}_k\|^2|\mathcal{F}_{k-1}] = \|\boldsymbol{\mu}\|^2 + d\sigma^2. \tag{59}$$

Combining (58) and (59), the result follows. ■

By Lemma 11 for each $k \ge 1$, we deduce

$$\mathbb{E}[V(\boldsymbol{\theta}_k)|\mathcal{F}_{k-1}] - V(\boldsymbol{\theta}_{k-1}) \le -\left(f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*)\right) + \frac{\alpha}{2}(\|\boldsymbol{\mu}\|^2 + d\sigma^2). \tag{60}$$

Therefore, in order to show that the pair (C,V) in (56) satisfies the drift equation (26), 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

$$f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) = \underbrace{f(\boldsymbol{\theta}_{k-1}) - f(\rho_{k-1}\boldsymbol{\mu})}_{(a)} + \underbrace{f(\rho_{k-1}\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*)}_{(b)}. \tag{61}$$

The assumption $\boldsymbol{\theta}_{k-1} \notin C$ yields that either $\sigma\|\tilde{\boldsymbol{\theta}}_{k-1}\| \ge c'$ or $|\rho_{k-1} - \rho^*| \ge \frac{1}{2}\rho^*$. In Lemma 12 (resp. 14), we show that (a) (resp. (b)) in (61) are both non-negative and they are lower bounded by some positive constant provided that $\sigma\|\tilde{\boldsymbol{\theta}}_{k-1}\| \ge c'$ and $|\rho_{k-1} - \rho^*| \le \frac{1}{2}\rho^*$ (resp. $|\rho_{k-1} - \rho^*| \ge \frac{1}{2}\rho^*$).

Lemma 12 *(Lower bound for (a) in (61)) 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. $f(\boldsymbol{\theta}) - f(\rho\boldsymbol{\mu}) \ge 0$.
2. $f(\boldsymbol{\theta}) - f(\rho\boldsymbol{\mu}) \ge 1$ *provided that* $|\rho - \rho^*| \le \frac{1}{2}\rho^*$, $\sigma\|\tilde{\boldsymbol{\theta}}\| \ge c'$ *and* $\sigma \ge c\|\boldsymbol{\mu}\|$ *where c is defined in* (39) *and* (40)*. Here ρ^* is defined in Lemma 2 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.

1. **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 (3). 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

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

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

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

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

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

Therefore, we have

$$\begin{aligned} 2\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\frac{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})}{1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right] &= 2\mathbb{E}_{\boldsymbol{\xi}}\left[\log(1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi}))\right] - \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})\right)^2\right] \\ &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1 + \frac{4\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}{(1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}))^2}\right)\right] \geq 0. \end{aligned} \tag{62}$$

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 $1 + \exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}) = 2\exp(-\frac{\rho\boldsymbol{\mu}^T\boldsymbol{\xi}}{2})\cosh(\frac{\rho\boldsymbol{\mu}^T\boldsymbol{\xi}}{2})$. 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 (62), we obtain that

$$\begin{aligned} 2\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\frac{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})}{1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right] &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1 + \frac{4\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}{(1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}))^2}\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1 + \frac{\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})}{\cosh^2(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right] \\ &\geq \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(1 + \frac{\sinh^2(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})}{\cosh^2(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\cdot 1_{\{\boldsymbol{\xi}:|\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}|\geq r\}}\right] \qquad (63) \\ &\geq \mathbb{E}_{\boldsymbol{\xi}}\left[\left(\log 2 + \log\left(\frac{|\sinh(\frac{\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}}{2})|}{\cosh(\frac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right)\cdot 1_{\{\boldsymbol{\xi}:|\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi}|\geq r\}}\right]. \end{aligned}$$

Here (63) 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), we have $\boldsymbol{\mu}^T\boldsymbol{\xi} \sim N(\|\boldsymbol{\mu}\|^2, \sigma^2\|\boldsymbol{\mu}\|^2)$ and $\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi} \sim N(0, \sigma^2\|\tilde{\boldsymbol{\theta}}\|^2)$, so $\tilde{\boldsymbol{\theta}}^T\boldsymbol{\xi} = \sigma\|\tilde{\boldsymbol{\theta}}\|\psi$ 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

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\frac{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{\xi})}{1+\exp(-\rho\boldsymbol{\mu}^T\boldsymbol{\xi})}\right)\right] \\ &\geq \tfrac{1}{2}\log(2)\cdot\mathbb{E}_{\psi}\left[1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] + \tfrac{1}{2}\mathbb{E}_{\psi}\left[\log\left|\sinh(\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{2})\right|1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] - \tfrac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[\log(\cosh(\tfrac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi}))\right] \\ &\geq \tfrac{1}{2}\log(2)\cdot\mathbb{E}_{\psi}\left[1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] + \tfrac{1}{2}\mathbb{E}_{\psi}\left[\log\left|\sinh(\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{2})\right|1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] - \tfrac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[|\tfrac{\rho}{2}\boldsymbol{\mu}^T\boldsymbol{\xi}|\right] \\ &\geq \tfrac{1}{2}\log(2)\cdot\mathbb{E}_{\psi}\left[1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] + \tfrac{1}{2}\mathbb{E}_{\psi}\left[\log\left|\sinh(\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{2})\right|1_{\{|\psi|\geq\frac{r}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] - \frac{3}{4}\left(\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2} + \sqrt{\frac{2}{\pi}}\cdot\frac{\|\boldsymbol{\mu}\|}{\sigma}\right), \end{aligned} \tag{64}$$

where the last inequality uses (5) 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

$$\begin{aligned}
&\tfrac{1}{2}\log(2) \cdot \mathbb{E}_\psi \left[1_{\{|\psi| \geq \frac{4\log(\sqrt{2}+1)\}}{\sigma\|\tilde{\boldsymbol{\theta}}\|}}\right] + \tfrac{1}{2}\mathbb{E}_\psi \left[\log \left|\sinh(\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{2})\right| 1_{\{|\psi| \geq \frac{4\log(\sqrt{2}+1)}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] \\
&\qquad \geq \tfrac{1}{2}\log(2) \cdot \mathbb{E}_\psi \left[1_{\{|\psi| \geq \frac{4\log(\sqrt{2}+1)}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] + \tfrac{1}{2}\mathbb{E}_\psi \left[\left|\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{4}\right| 1_{\{|\psi| \geq \frac{4\log(\sqrt{2}+1)}{\sigma\|\tilde{\boldsymbol{\theta}}\|}\}}\right] \\
&\qquad \geq \tfrac{1}{2}\log(2) \cdot \mathbb{E}_\psi [1_{\{|\psi| \geq 1\}}] + \tfrac{1}{2}\mathbb{E}_\psi \left[\left|\tfrac{\sigma\|\tilde{\boldsymbol{\theta}}\|\psi}{4}\right| 1_{|\psi| \geq 1\}}\right] \qquad (65) \\
&\qquad \geq \left(\tfrac{1}{2}\log(2) + \frac{\sigma\|\tilde{\boldsymbol{\theta}}\|}{8}\right) \cdot \Phi^c(1).
\end{aligned}$$

Here (65) follows from the assumption that $\sigma\|\tilde{\boldsymbol{\theta}}\| \geq 436$. Combining (64), (65) and the bounds $\sigma \geq 0.33\|\boldsymbol{\mu}\|$ and $\sigma\|\tilde{\boldsymbol{\theta}}\| \geq 436$ the result follows.

2. **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

$$\kappa(\xi_1, \xi_2) = 0 \text{ whenever } |\xi_1| \leq |1 - \rho\xi_2|. \qquad (66)$$

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 (66) is established. We claim the following

$$\kappa(\xi_1, \xi_2) = |\xi_1| - |1 - \rho\xi_2| \text{ whenever } |\xi_1| \geq |1 - \rho\xi_2|. \qquad (67)$$

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 (67) follows from the following simple identity

$$2\ell(t) = 1 - t + |1 - t|, \quad \forall t \in \boldsymbol{R}. \qquad (68)$$

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 (67) follows from the identity (68). We therefore obtain

$$\begin{aligned} \mathbb{E}_{\xi_1,\xi_2}[\kappa(\xi_1,\xi_2)] &= 2\mathbb{E}_{\xi_1,\xi_2}[(\ell(\xi_1+\rho\xi_2)+\ell(-\xi_1+\rho\xi_2)-2\ell(\rho\xi_2))1_{\{\xi_1>0\}}] \quad (69)\\ &= 2\mathbb{E}_{\xi_1,\xi_2}[(\ell(\xi_1+\rho\xi_2)+\ell(-\xi_1+\rho\xi_2)-2\ell(\rho\xi_2))1_{\{\xi_1\geq|1-\rho\xi_2|\}}] \quad (70)\\ &= \mathbb{E}_{\xi_1,\xi_2}[(\xi_1-|1-\rho\xi_2|)1_{\{\xi_1\geq|1-\rho\xi_2|\}}]. \quad (71) \end{aligned}$$

Here equation (69) holds because $\xi_1 \sim -\xi_1$ and $\kappa(\xi_1,\xi_2) = \kappa(-\xi_1,\xi_2)$. Equation (70) is true because of claim (66) and (71) follows from claim (67). From (71), we conclude that $\mathbb{E}_{\xi_1,\xi_2}[\kappa(\xi_1,\xi_2)] \geq 0$. We then observe the bound

$$\begin{aligned} \mathbb{E}_{\xi_1,\xi_2}[(\xi_1-|1-\rho\xi_2|)1_{\{\xi_1\geq|1-\rho\xi_2|\}}] &= \tfrac{1}{2}\mathbb{E}_{\xi_1,\xi_2}[\xi_1-|1-\rho\xi_2|+|\xi_1-|1-\rho\xi_2||]\\ &\geq -\tfrac{1}{2}\mathbb{E}_{\xi_2}[|1-\rho\xi_2|]+\tfrac{1}{2}\mathbb{E}_{\xi_1,\xi_2}[|\xi_1|-|1-\rho\xi_2|] \quad (72)\\ &= \tfrac{1}{2}\mathbb{E}_{\xi_1}[|\xi_1|]-\mathbb{E}_{\xi_2}[|1-\rho\xi_2|]. \end{aligned}$$

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

$$\mathbb{E}_{\xi_1}[|\xi_1|] = \sqrt{\frac{2}{\pi}}\cdot\sigma\|\tilde{\boldsymbol{\theta}}\|, \tag{73}$$

and

$$\mathbb{E}_{\boldsymbol{\xi}}[|1-\rho\boldsymbol{\mu}^T\boldsymbol{\xi}|] \leq 1+\rho\mathbb{E}_{\boldsymbol{\xi}}[|\boldsymbol{\mu}^T\boldsymbol{\xi}|] \leq 1+\rho\|\boldsymbol{\mu}\|\left(\sqrt{\frac{2}{\pi}}\cdot\sigma+\|\boldsymbol{\mu}\|\right). \tag{74}$$

Combing equations (66), (67), (72), (73), and (74), we deduce

$$f(\boldsymbol{\theta})-f(\rho\boldsymbol{\mu}) \geq \frac{1}{2}\left(\sqrt{\frac{1}{2\pi}}\cdot\sigma\|\tilde{\boldsymbol{\theta}}\|-1-\rho\|\boldsymbol{\mu}\|\left(\sqrt{\frac{2}{\pi}}\cdot\sigma+\|\boldsymbol{\mu}\|\right)\right). \tag{75}$$

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 (75).

■

We next derive a lower bound (61), Part (b). But, first we need a basic lemma from convex analysis.

Lemma 13 *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*

$$g(\rho)-g(\rho^*) \geq \frac{\rho^* B}{8}|\rho-\rho^*| \quad \textit{for all } \rho \notin [\tfrac{1}{2}\rho^*, \tfrac{3}{2}\rho^*]. \tag{76}$$

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

Lemma 14 *(Lower bound for (b) in (61)) 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 (39) and (40), there exists a positive constant A such that the following is true*

$$f(\rho\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*) \geq A \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}. \tag{77}$$

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 2, 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 (77), we instead aim to bound this difference in the function g. From (2), 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, for $\rho \in [\frac{3}{4}\rho^*, \frac{5}{4}\rho^*]$, the term $\exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2} - \frac{1}{2} - \left|\frac{\|\boldsymbol{\mu}\|}{\sigma} - \rho\sigma\|\boldsymbol{\mu}\|\right|\right)$ is lower bounded by $\exp\left(-\frac{1}{2c^2} - \frac{1}{4c} - \frac{1}{2}\right)$. By Lemma 13, 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)*, 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*

$$h'(\rho) = \frac{-1}{\sigma\|\boldsymbol{\mu}\|\sqrt{2\pi}} \int_{-\infty}^{\frac{1}{\rho}} z \exp\left(-\frac{1}{2} \cdot \left(\frac{z}{\sigma\|\boldsymbol{\mu}\|} - \frac{\|\boldsymbol{\mu}\|}{\sigma}\right)^2\right) dz. \tag{78}$$

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 13 and* $|\rho - \rho^*| \geq \frac{1}{2}\rho^*$*, it holds that*

$$f(\rho\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*) \geq \frac{4}{125\sqrt{2\pi}} \cdot \frac{\sigma}{r\|\boldsymbol{\mu}\|} \cdot \exp\left(-\frac{1}{2} \cdot \Gamma^2\right). \tag{79}$$

Here $r = \rho^* \sigma^2$*. Note that* $r > 0$ *by Lemma 2. We aim to lower bound the right-hand side of* (79)*. We denote by* $w = \frac{\sigma}{r\|\boldsymbol{\mu}\|} - \frac{\|\boldsymbol{\mu}\|}{\sigma}$ *the quantity defined in Lemma 2. In particular, by Lemma 2, the following holds*

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{\sigma}{\|\boldsymbol{\mu}\|} = \Phi(w) \cdot \exp(\tfrac{1}{2}w^2). \tag{80}$$

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* $w^2 \geq \frac{1}{2} \cdot \Gamma^2$ *for all* $w \geq \frac{1}{3\sqrt{2}-4} \cdot \frac{\|\boldsymbol{\mu}\|}{\sigma}$*. On the other hand, by* (80) *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* $w^2 \geq \frac{1}{2} \cdot \Gamma^2$, $\exp(-w^2) \geq \frac{\pi}{2} \cdot \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$*, and* $\frac{\sigma}{r\|\boldsymbol{\mu}\|} \geq w \geq \frac{1}{(3\sqrt{2}-4)c}$ *into the right-hand-side of* (79)*, 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* (79) *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* (79)*. 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 (80) *is an increasing function with respect to w. Therefore it suffices to show that the following holds*

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{\sigma}{\|\boldsymbol{\mu}\|} \geq \Phi\left(-\frac{\|\boldsymbol{\mu}\|}{2\sigma}\right) \cdot \exp\left(\frac{\|\boldsymbol{\mu}\|^2}{8\sigma^2}\right) \quad \textit{whenever} \quad \frac{\sigma}{\|\boldsymbol{\mu}\|} \geq c. \tag{81}$$

However, it can be verified by a plot that $\frac{1}{\sqrt{2\pi}} \geq t \cdot \Phi\left(-\frac{t}{2}\right) \cdot \exp\left(\frac{t^2}{8}\right)$ 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 (77) *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 6.

Proof [Proof of Theorem 6] Consider the set C and function V defined in (56):

$$C := \left\{\boldsymbol{\theta} : |\rho - \rho^*| < \tfrac{1}{2}\rho^* \text{ and } \sigma\|\tilde{\boldsymbol{\theta}}\| \leq c'\right\} \quad \text{and} \quad V(\boldsymbol{\theta}) = \frac{1}{2\alpha}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2. \tag{82}$$

We let c' to be defined as in Lemma 12. 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

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} \geq 1\right) \geq \delta \quad \text{for all} \quad \boldsymbol{\theta} \in C. \tag{83}$$

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 (3), we see that $\boldsymbol{\xi}^T\boldsymbol{\theta}$ and $\boldsymbol{\xi}^T\tilde{\boldsymbol{\theta}}$ are independent normal random variables. It thus holds that

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} \geq 1\right) \geq \mathbb{P}_{\boldsymbol{\xi}}\left(\rho\boldsymbol{\xi}^T\boldsymbol{\mu} \geq 1\right) \cdot \mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\tilde{\boldsymbol{\theta}} \geq 0\right) = \frac{1}{2} \cdot \mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\boldsymbol{\mu} \geq \frac{1}{\rho}\right). \tag{84}$$

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

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} \geq 1\right) \geq \delta := \frac{1}{2} \cdot \Phi^c\left(\frac{\frac{2}{\rho^*} - \|\boldsymbol{\mu}\|^2}{\sigma\|\boldsymbol{\mu}\|}\right). \tag{85}$$

We next show that the pair (C, V) satisfies the drift equation (26). Let us rewrite (61):

$$f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) = \underbrace{f(\boldsymbol{\theta}_{k-1}) - f(\rho_{k-1}\boldsymbol{\mu})}_{(a)} + \underbrace{f(\rho_{k-1}\boldsymbol{\mu}) - f(\boldsymbol{\theta}^*)}_{(b)}. \tag{86}$$

By Lemmas 12 and 14, both terms in (a) and (b) in (86) 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 14 satisfies $1 \geq A\cdot\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$ for all $\frac{\sigma}{\|\boldsymbol{\mu}\|} \geq c$. Hence it holds that

$$A\cdot\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2} \leq f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) \quad \text{for all } \boldsymbol{\theta}_{k-1} \notin C. \tag{87}$$

We use (57) next to establish the drift equation (26). Recall that the following holds

$$f(\boldsymbol{\theta}_{k-1}) - f(\boldsymbol{\theta}^*) \leq \frac{1}{2\alpha}\left(\|\boldsymbol{\theta}_{k-1}-\boldsymbol{\theta}^*\|^2 - \mathbb{E}\left[\|\boldsymbol{\theta}_k-\boldsymbol{\theta}^*\|^2 \,|\mathcal{F}_{k-1}\right]\right) + \frac{\alpha}{2}\left(\|\boldsymbol{\mu}\|^2 + d\sigma^2\right). \tag{88}$$

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

$$\left(\mathbb{E}\left[V(\boldsymbol{\theta}_k)|\mathcal{F}_{k-1}\right] - V(\boldsymbol{\theta}_{k-1})\right)\cdot 1_{\{\boldsymbol{\theta}_{k-1}\notin C\}} \leq \left(\frac{\alpha}{2}(\|\boldsymbol{\mu}\|^2+d\sigma^2) - A\cdot\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}\right)\cdot 1_{\{\boldsymbol{\theta}_{k-1}\notin C\}}. \tag{89}$$

Therefore, by choosing $\alpha < A\cdot\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2(\|\boldsymbol{\mu}\|^2+d\sigma^2)}$, we obtain the drift equation (26) 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 9 and a simple induction, we obtain that

$$\mathbb{E}[\tau_m] \leq \tfrac{1}{b}V(0) + \tfrac{1}{b}(m-1)\sup_{\boldsymbol{\theta}\in C} V(\boldsymbol{\theta}). \tag{90}$$

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

$$\mathbb{E}[\tau_m] \leq \gamma\cdot m. \tag{91}$$

Combining (91), (85) and Lemma 4, the proof immediately follows. ■

4.3. Angle bound, proof of Theorem 7

Proof [Proof of Theorem 7] 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

$$\left||\boldsymbol{v}^T\boldsymbol{\theta}_k| - |\boldsymbol{v}^T\boldsymbol{\theta}_{k-1}|\right| \leq \alpha|\boldsymbol{v}^T\boldsymbol{\xi}_{k-1}|. \tag{92}$$

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 $k \geq 1$

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

Here we used that $\boldsymbol{v}^T\boldsymbol{\xi}_k \sim N(0,\sigma^2)$ along with (5). We also see that

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

Therefore, we have shown that $X_0, X_1, \cdots$ is a super-martingale. By Theorem 1, we have $\mathbb{E}[X_T] \leq 0$. The result follows.

■

5. Numerical Experiments

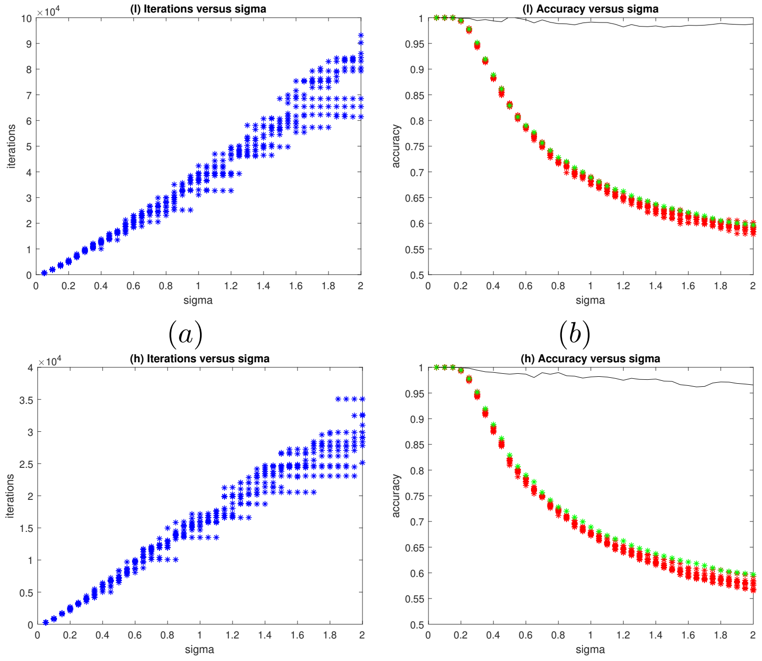
We investigate the performance of our termination test on two popular data sets, MNIST [15] and CIFAR-10 [13], as well as synthetic data generated from Gaussians and heavy-tailed student t-distributions. All tests were performed using our zero overhead stopping criteria outlined in (12); experiments using our test which required an extra sample (11) 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) [14]. The SVS termination test is as follows. One fixes a validation set of p instances $(\boldsymbol{\zeta}_1^{\text{V}}, y_1^{\text{V}}), \ldots, (\boldsymbol{\zeta}_p^{\text{V}}, y_p^{\text{V}})$ 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 (12) is 0. The value of p is a tuning parameter for SVS; we exhibit results for three different p values (see Figs. 2, 3, 4, 5).

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 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.



(a) (b)

(c) (d)

Figure 1: Performance of stopping criterion (12) 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 (12) 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 (12) holds to the accuracy of the optimal classifier.

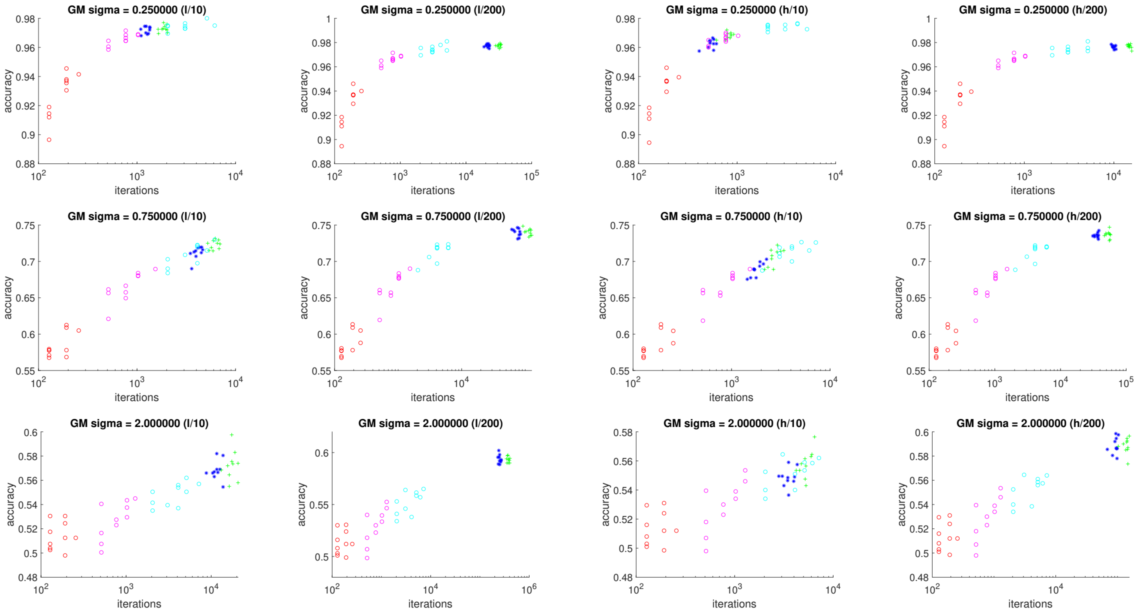


Figure 2: Each plot shows 10 random runs of SGD applied to normally distributed data with indicated values of σ and for a fixed dimension $d = 500$. For each of the ten runs, five termination tests corresponding to five colors were applied. SVS was tried with $p = 32, 128, 512$, depicted as red, magenta and cyan circles respectively. Test (12) is indicated with a blue asterisk. A green '+' corresponds to termination after $1.5k$ iterations, where k is the iteration index that (12) 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.

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., [1]) 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 $\mathbb{E}\left[\|\boldsymbol{\theta}_k\|\right] \leq O(\alpha k)$. See also Theorems 5 and 6. 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.

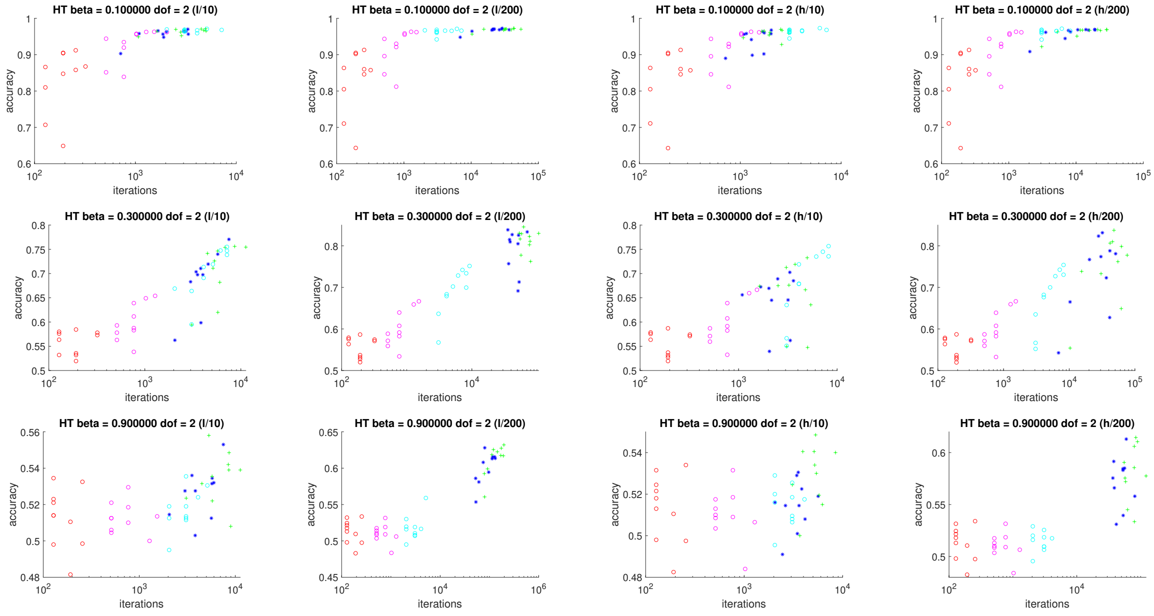


Figure 3: Tests on the student-t distribution (heavy tailed) with two degrees of freedom and the indicated value of parameter β. See the caption of Fig. 2 for explanation of the plots.

5.1. 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. 1, 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 (12) holds almost matches the optimal classifier; in particular, the averaged accuracy of our classifier/accuracy of the optimal classifier over the 10 runs, black curve in Fig. 5, never dips below 0.95.

In Fig. 2, we compare performance of (12) 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 (12) 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. 2 indicate that (12) 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 $\tilde{\alpha}$, (12) 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

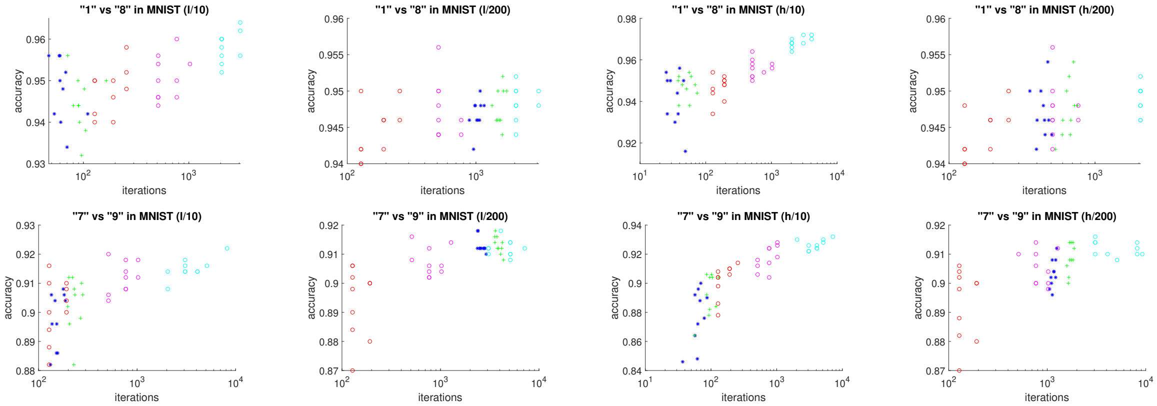


Figure 4: Tests on the MNIST handwritten digit data set for discerning "1" from "8" and "7" from "9" for both hinge and logistic, and for both $\tilde{\alpha} = 1/10$ and $\tilde{\alpha} = 1/200$. Refer to the caption of Fig. 2 for the key to the plots.

entry is incremented by 1. Fig. 3 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 (12) is more spread out presumably because of the heavy-tailed nature of the data set.

5.2. Experiments with real data

MNIST handwritten digits. We compared our termination test on the MNIST handwritten digit set [15] ($d = 784$, 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. 4 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 [13] ($d = 3072$, 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.

6. Conclusions

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. The bounds that we proved are weaker than what we observed in our experiments. Therefore, the first obvious question left open by this work is whether the theoretical bounds can be improved.

In our experimental results, the plots in Figs. 2 through 5 show a consistent pattern that (12) 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

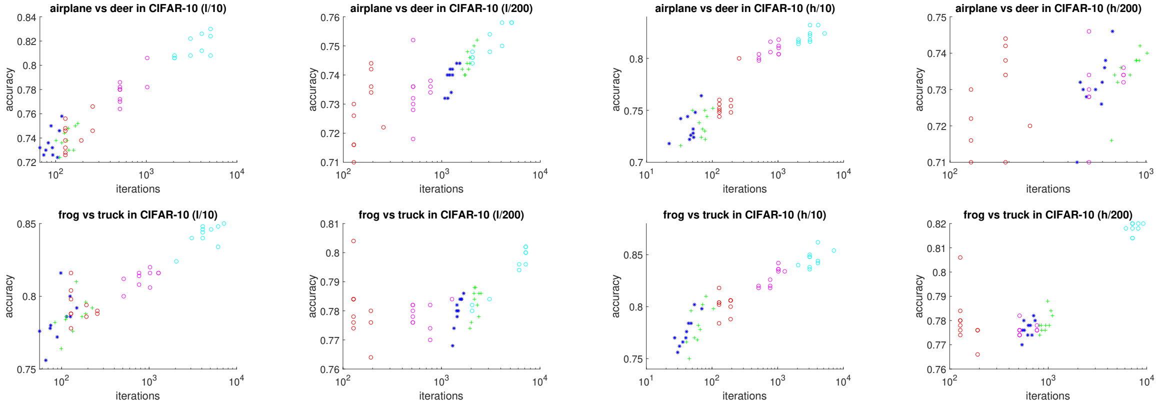


Figure 5: Tests on the CIFAR-10 image set for two tasks, for logistic and hinge losses, and for $\tilde{\alpha} = 1/10$ and $\tilde{\alpha} = 1/200$. Refer to the caption of Fig. 2 for the key to the plots. The plot in the first row, right, does not include cyan circles because the training data was exhausted before the SVS test could activate for $p = 512$.

the accuracy/iterations tradeoff in the hands of the user who selects the stepsize $\tilde{\alpha}$. Another benefit of (12) 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 subproblem of a larger computation.

This work did not explore regularization via early stopping. As mentioned in the introduction, experiments showed that as SGD iterations continued, the accuracy on the test set eventually levels off but does not decrease significantly, *i.e.*, SGD for binary classification is not prone to overfitting. Because the test accuracy never shows marked decline, there is no opportunity for early stopping to regularize. However, we know of other settings in which early stopping has a strong regularizing effect (*e.g.*, conjugate gradient iterations for image deconvolution, already known in [30]), so if (12) is extended beyond binary classification in future work, there will likely also be an opportunity to explore regularization.

References

[1] Nemirovski A., Juditsky A., Lan G., and Shapiro A. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

[2] H. Ashtiani, S. Ben-David, N. J. A. Harvey, C. Liaw, A. Mehrabian, and Y. Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In *Advances in Neural Information Processing Systems (NeurIPs)*, 2018.

[3] Juditsky A. B., Nazin A. V., Nemirovsky A. S., and Tsybakov A. B. Algorithms of robust stochastic optimization based on mirror descent method. *preprint arXiv:1907.02707*, 2019.

[4] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[5] S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8 (3-4):231–357, 2015.

[6] Drusvyatskiy D. and Davis D. Robust stochastic optimization with the proximal point method. *preprint arXiv:1907.13307*, 2019.

[7] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, New York, NY, USA, 4th edition, 2010.

[8] D. Duvenaud, D. Maclaurin, and R. P. Adams. Early Stopping as Nonparametric Variational Inference. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[9] F. Famoye. Continuous univariate distributions, volume 1. *Technometrics*, 37:466–466, 11 1995.

[10] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2016.

[11] Ziwei J. and Telgarsky M. Risk and parameter convergence of logistic regression. *preprint arXiv:1803.07300*, 2018.

[12] T. Jiang, S. A. Vavasis, and C. W. Zhai. Recovery of a mixture of gaussians by sum-of-norms clustering. *preprint arXiv:1902.07137*, 2019.

[13] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[14] Prechelt L. *Early Stopping — But When?*, pages 53–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[15] 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.

[16] J. Lin and L. Rosasco. Optimal learning for multi-pass stochstic gradient methods. In *Advances in Neural Information Processing Systems (NeurIPs)*, pages 4556–4564, 2016.

[17] 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 (ICML)*, pages 2340–2348, 2016.

[18] M. Lui and C. Guang. Early stopping for nonparametric testing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[19] Nacson M., Srebro N., and Soudry D. Stochastic Gradient Descent on Separable Data. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[20] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[21] D. Molitor, D. Needell, and R. Ward. Bias of homotopic gradient descent for the hinge loss. *preprint arXiv:1907.11746*, 2019.

[22] A. Panahi, D. Dubhashi, F. D. Johansson, and C. Bhattacharyya. Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2769–2777, 2017.

[23] G. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM J. Control Optim.*, 24(4):655–666, 1986.

[24] 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.

[25] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[26] Ghadimi S. and Lan G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, i: a generic algorithmic framework. *SIAM J. Optim.*, 22 (4):1469–1492, 2012.

[27] Ghadimi S. and Lan G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: Shrinking procedures and optimal algorithms. *SIAM J. Optim.*, 23(4):2061–2089, 2013.

[28] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[29] S. Sherman and T. G. Kolda. Estimating higher-order moments using symmetric tensor decomposition. *preprint arXiv:1911.03813*, 2019.

[30] A. van der Sluis and H. van der Vorst. SIRT-and CG-type methods for the iterative solution of sparse linear least-squares problems. *Linear Algebra Appl.*, 130:257–303, 1990.

[31] Yao Y., Rosasco L., and Caponnetto A. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.