Implicit Bias of Gradient Descent on Separable Data

Ziniu Li
ziniuli@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

March 9, 2021

Mainly based on the paper:
Soudry, Daniel, et al. "The implicit bias of gradient descent on separable data." The Journal of Machine Learning Research 19.1 (2018): 2822-2878.

Outline

Outline

Minimum ℓ_2-norm Solution to Undetermined Linear System

- Consider to solve a undetermined linear system:

$$Ax = b,$$

  where $A \in \mathbb{R}^{m \times n}$ with full row rank $(m < n)$ and $b \in \mathbb{R}^m$.

- There are infinite solutions since A has full row rank.
- If we consider to seek the minimum ℓ_2-norm solution, the optimal solution is unique:

$$\min_x \|x\|, \quad \text{s.t. } Ax = b. \tag{1}$$

- The optimal solution x^* to Eq. 1 is given by

$$x^* = A^\top (AA^\top)^{-1} b. \tag{2}$$

  – see for example http://www.math.usm.edu/lambers/mat419/lecture15.pdf for the derivation.

Minimum ℓ_2-norm Solution to Undetermined Linear System

- To facilitate later analysis, let's consider the singular value decomposition for A:

$$A = U\Sigma V^\top = U \begin{bmatrix} \Sigma_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix} = U\Sigma_1 V_1^\top,$$

  where $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ contains singular values ($\Sigma_1 \in \mathbb{R}^{m\times m}, V_1 \in \mathbb{R}^{n\times m}$).

- We can show that the optimal solution in Eq. 2 is

$$\begin{aligned} x^* &= A^\top (AA^\top)^{-1} b \\ &= V_1 \Sigma_1 U^\top \left(U\Sigma_1 V_1^\top V_1 \Sigma_1 U^\top\right)^{-1} b \\ &= V_1 \Sigma_1^{-1} U^\top b. \end{aligned} \tag{3}$$

Gradient Descent For the Least Square Problem

Question: If we consider the gradient descent (GD) on the least square problem, which solution does GD converge to?

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} \|Ax - b\|^2, \tag{4}$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ follow the same condition as before (i.e., A has full row rank).

Gradient Descent Finds the Minimum ℓ_2-norm Solution

Specifically, we have the following update rule:

$$x_{k+1} = x_k - \eta \nabla f(x_k) = x_k - \eta A^\top (Ax_k - b), \tag{5}$$

where $\eta > 0$ is the stepsize (learning rate). To better study the asymptotic behavior (the limit of x_k when $k \to \infty$), let's rewrite Eq. 5 as

$$\begin{aligned} x_k &= \left(I - \eta A^\top A\right) x_{k-1} + \eta A^\top b \\ &= \cdots \\ &= \left(I - \eta A^\top A\right)^k x_0 + \eta \sum_{\ell=0}^{k-1} \left(I - \eta A^\top A\right)^\ell A^\top b. \end{aligned}$$

Gradient Descent Finds the Minimum ℓ_2-norm Solution

Let $y_k = V^\top x_k \in \mathbb{R}^m$, we have that

$$\begin{aligned} y_k &= \left(I - \eta\Sigma^\top\Sigma\right)^k y_0 + \eta \sum_{\ell=0}^{k-1} \left(I - \eta\Sigma^\top\Sigma\right)^\ell \Sigma^\top U^\top b \\ &= \begin{bmatrix} \left(I - \eta\Sigma_1^2\right)^k & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} y_0 + \eta \sum_{\ell=0}^{k-1} \begin{bmatrix} \left(I - \eta\Sigma_1^2\right)^\ell & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ \mathbf{0} \end{bmatrix} U^\top b \\ &= \begin{bmatrix} \left(I - \eta\Sigma_1^2\right)^k & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} y_0 + \eta \sum_{\ell=0}^{k-1} \begin{bmatrix} \left(I - \eta\Sigma_1^2\right)^\ell \Sigma_1 \\ \mathbf{0} \end{bmatrix} U^\top b. \end{aligned}$$

Therefore, we see that as long as η is small such that $I - \eta\Sigma_1^2 \succeq 0$ (that is, $\eta < \sigma_{\max}^{-2}(A)$, where $\sigma_{\max}$ is the maximal singular value of A), then $y_k \to y_\infty$.

Gradient Descent Finds the Minimum ℓ_2-norm Solution

$$
\begin{aligned}
y_\infty &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} y_0 + \eta \sum_{\ell=0}^{\infty} \begin{bmatrix} \left(I - \eta \Sigma_1^2\right)^\ell \Sigma_1 \\ \mathbf{0} \end{bmatrix} U^\top b \\
&= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} V^\top x_0 + \eta \begin{bmatrix} \left(I - \left(I - \eta \Sigma_1^2\right)\right)^{-1} \Sigma_1 \\ \mathbf{0} \end{bmatrix} U^\top b \\
&= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix} x_0 + \eta \Sigma_1^{-1} U^\top b \\
&= V_2^\top x_0 + \eta \Sigma_1^{-1} U^\top b.
\end{aligned}
$$

Hence, $x_\infty = V_2 y_\infty = V_2 V_2^\top x_0 + V_1 \Sigma_1^{-1} U^\top b$. Therefore, if x_0 is orthogonal to the null space of A, then the gradient descent will converge to the minimum ℓ_2-norm solution.

Classification

Observation: GD finds the ℓ_2-norm solution for undetermined least square (regression) problem.

Question: Does the same magic hold for classification problem?

Answer: Yes! GD finds the max margin solution for linearly separable binary classification problem without an explicit regularization!

Outline

Problem Setting

- Consider a dataset $\{x_n, y_n\}_{n=1}^N$ with $x_n \in \mathbb{R}^d$ and $y_n \in \{-1, 1\}$.
- The empirical loss functions is:

$$L(w) = \sum_{n=1}^{N} \ell\left(y_n w^\top x_n\right), \tag{6}$$

  where $w \in \mathbb{R}^d$ is the weight vector.
  - Decision rule: predict 1 if $w^\top x_n \geq 0$ and -1 otherwise.
- To simplify notation, we augment y_n into x_n so that the loss function becomes:

$$L(w) = \sum_{n=1}^{N} \ell\left(w_n^\top x_n\right).$$

Main Assumptions

Assumption 1.

The dataset is linearly separable: $\exists w_*$ *s.t.* $\forall n : w_*^\top x_n > 0$.

Assumption 2.

The loss function $\ell(u)$ *is positive, differentiable, monotonically decreasing to zero, that is,* $\forall u : \ell(u) > 0, \ell'(u) < 0, \lim_{u\to\infty} \ell(u) = \lim_{u\to\infty} \ell'(u) = 0$. *Moreover, it is a* β*-smooth function, i.e., its derivative is* β*-Lipschitz continuous and* $\lim_{u\to-\infty} \ell'(u) \neq 0$.

Main Assumptions

- Assumption 2 includes many common loss functions:
  - logistic: $\ell(u) = \log(1 + \exp(-u))$.
  - exp-loss: $\ell(u) = \exp(-u)$.
- Assumption 2 implies that $L(w)$ is $\beta\sigma_{\max}^2(X)$-smooth function, where $\sigma_{\max}(X)$ is the maximal singular value of $X \in \mathbb{R}^{d\times N}$.
- The above assumptions imply that the infimum of the optimization problem is zero, but it is not attached at any finite w. Moreover, no finite critical point w exists.

Global Convergence

Let's consider the gradient descent method with a fixed stepsize η:

$$w(t+1) = w(t) - \eta \nabla L\left(w(t)\right) = w(t) - \eta \sum_{n=1}^{N} \ell'\left(w(t)^\top x_n\right) x_n. \tag{7}$$

Lemma 1.

Let $w(t)$ be the iterates of gradient descent (Eq. 7) with $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(X)$ and any starting point $w(0)$. Under Assumption 1 and 2, we have

(1) $\lim_{t\to\infty} L\left(w(t)\right) = 0.$

(2) $\lim_{t\to\infty} \|w(t)\| = 0.$

(3) $\forall n : \lim_{t\to\infty} w(t)^\top x_n = \infty.$

Proof of Lemma 1

Note that data is linearly separable, which means that $\exists w_*$ such that

$$w_*^\top \nabla L(w) = \sum_{n=1}^{N} \underbrace{\ell'(w^\top x_n)}_{<0} \underbrace{w_*^\top x_n}_{>0}.$$

- For any finite w, the above sum cannot be equal to zero. Therefore, no finite critical points w for which $\nabla L(w) = 0$.
- Note that gradient descent on a smooth loss function with a small stepsize is always guaranteed to converge to a critical point: $\nabla L(w(t)) \to 0$.
- Therefore, $\|w(t)\| \to \infty$, which also implies $L(w) \to 0$. So GD converges to the global minimum.

Implication of Lemma 1

- It's meaningless to study $\|w\|$ since $\|w(t)\| \to \infty$.
- Instead, we should study the direction of $w(t)$.
- That is, if the limit $\lim_{t\to\infty} w(t)/\|w(t)\|$ exists, what is it?

More Assumption

Definition 1.

A function $f(u)$ has a "tight exponential tail", if there exist positive constants, c, α, μ_+, μ_- such that

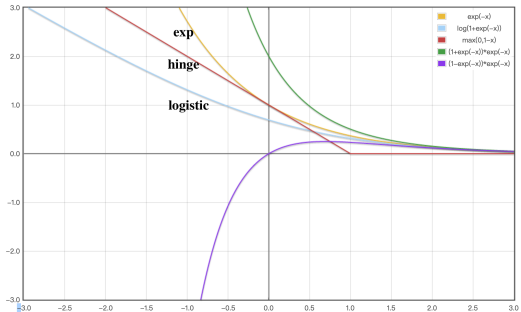
$$\forall u > u_+ : \quad f(u) \le c\,(1 + \exp(-u_+ u))\exp(-au),$$
$$\forall u > u_- : \quad f(u) \ge c\,(1 - \exp(-u_- u))\exp(-au).$$

Assumption 3.

The negative loss derivative $-\ell'(u)$ has a tight exponential tail.

More Assumption

For example, the exponential loss $\ell(u) = \exp(-u)$ and the logistic loss $\ell(u) = \log(1 + \exp(-u))$ follow this assumption with $a = c = 1$.



Without loss of generality, we assume $a = c = 1$ since these constants can be normalized by rescaling x_n.

Outline

Outline

Main Result

Theorem 1.

For any dataset which is linearly separable (see Assumption 1), any β-smooth decreasing loss function (see Assumption 2) with an exponential tail (see Assumption 3), any stepsize $\beta < 2\beta^{-1}\sigma_{\max}^{-2}(X)$ and any starting point $w(0)$, the gradient descent iterates will behave as:

$$w(t) = \widehat{w}\log(t) + \rho(t), \tag{8}$$

where $\widehat{w}$ is the L_2 max margin vector (the solution to the hard margin SVM):

$$\widehat{w} = \arg\min_{w\in\mathbb{R}^d} \|w\|^2 \quad \text{s.t.} \quad w^\top x_n \geq 1, \tag{9}$$

and the residual grows at most as $\|\rho(t)\| = \mathcal{O}(\log\log(t))$ and so

$$\lim_{t\to\infty} \frac{w(t)}{\|w(t)\|} = \frac{\widehat{w}}{\|\widehat{w}\|}.$$

Furthermore, for almost all datasets (except measure zero), the residual $\rho(t)$ is bounded.

Limit the Residual

Theorem 2.

Under the same conditions and notations of Theorem 1, for almost all datasets (except measure zero), if in addition the support vectors $\mathcal{S} = \arg\min_n \widehat{w}^\top x_n$ *span the data (i.e.,* $\mathbf{rank}(X_\mathcal{S}) = \mathbf{rank}(X)$*, where* $X_\mathcal{S}$ *is a matrix whose columns are only those data points* x_n *s.t.* $\widehat{w}^\top x_n = 1$*), then*

$$\lim_{t\to\infty} \rho(t) = \widetilde{w},$$

where $\widetilde{w}$ *is a solution to:*

$$\forall n \in \mathcal{S}: \quad \eta \exp\left(-x_n^\top \widetilde{w}\right) = \alpha_n. \tag{10}$$

Outline

Implication: Rates of convergence

Theorem 3.

Under the conditions and notations of Theorem 1, for any linearly separable data set, the normalized weight vector converges to the normalized max margin vector in ℓ_2-norm:

$$\left\| \frac{w(t)}{\|w(t)\|} - \frac{\widehat{w}}{\|\widehat{w}\|} \right\| = \mathcal{O}\left(\frac{\log\log t}{\log t} \right),$$

with this rate improving to $\mathcal{O}(1/\log t)$ for almost every dataset; and in angle

$$1 - \frac{w(t)^\top \widehat{w}}{\|w(t)\| \, \|\widehat{w}\|} = \mathcal{O}\left(\left(\frac{\log\log t}{\log t} \right)^2 \right),$$

with this rate improving to $\mathcal{O}(1/\log^2 t)$ for almost every dataset; and the margin converges as

$$\frac{1}{\|\widehat{w}\|} - \frac{\min_n x_n^\top w(t)}{\|w(t)\|} = \mathcal{O}\left(\frac{1}{\log t} \right).$$

On the other hand, the loss itself decreases as

$$L\left(w(t)\right) = \mathcal{O}\left(\frac{1}{t} \right).$$

Remark on Theorem 3

- It's possible to show that all rates in Theorem 3 are tight.
- The training loss decreases fast than the margin gap $\rightsquigarrow$
  - We need to wait until the loss is exponentially small (note that loss is always positive) in order to be close to the max margin solution.
  - Similar to Adaboost, this helps explain why continuing to optimize the training loss, even when the loss is extremely small or even zero, still improves the generalization performance.
- Somewhat surprisingly, the population loss may increase while the population (classification) error decreases over iterations (see the next page for more explanation).

Explanation on Population Loss

- Some facts that we need to remember
  - $w(t)$ converges to $\widehat{w}$ as $t \to \infty$.
  - $\widehat{w}$ has zero training loss but generally does not have zero population misclassification error.
- For example, consider the logistic loss $\ell(u) = \log\left(1 + \exp(-u)\right)$ and define the hinge-at-zero loss $h(u) = \max(0, -u)$.
  - We know that $\widehat{w}$ correctly classifies all training points, i.e., $\sum_{n=1}^{N} h(\widehat{w}^\top x_n) = 0$.
  - But, we expect that $\mathbb{E}\left[h(\widehat{w}^\top x)\right] > 0$.
- Since $w(t) \approx \widehat{w} \log t$ and $\ell(\alpha u) \to \alpha h(u)$ as $\alpha \to \infty$, so

$$\mathbb{E}\left[\ell\left(w(t)^\top x\right)\right] \approx \mathbb{E}\left[\ell\left(\log(t)\widehat{w}^\top x\right)\right] \approx \log(t)\mathbb{E}\left[h(\widehat{w}^\top x)\right] = \Omega(\log t). \tag{11}$$

- The population loss increases logarithmically!

Explanation on Population Loss

Corollary 1.

Let ℓ be the logistic loss and $\mathcal{V}$ be an independent validation set, for which $\exists x \in \mathcal{V}$ such that $x^\top \widehat{w} < 0$. Then the validation loss increases as

$$L_{\mathrm{val}}\left(w(t)\right) = \sum_{x \in \mathcal{V}} \ell\left(w(t)^\top x\right) = \Omega(\log t).$$

Implication: we should use validation classification error rather than validation loss as the monitor to manage early stopping.

Connection with Adaboost

Adaboost can be viewed as the optimization algorithm of coordinate descent (along with the line search) on the exp-loss [Zhang and Yu, 2005, Telgarsky, 2013].

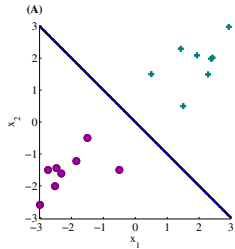
- An introductory material is https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/MIT15_097S12_lec10.pdf.

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} \exp(-y_i f(x_i)), \tag{12}$$

where $f(x_i) = \sum_{j=1}^{n} w_j h(x_i)$ is the output that integrates n weak learners and w is the weight for each learner.

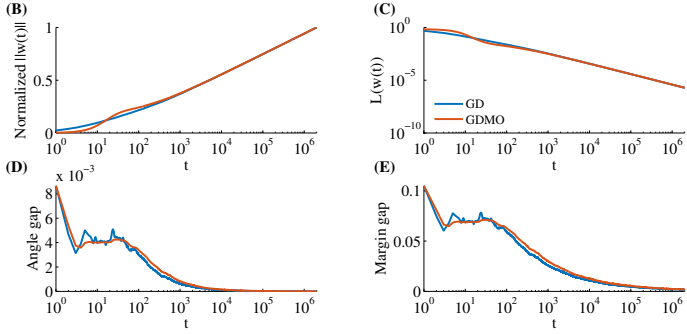
Outline

Artificial Dataset (Binary Classification)



Support vectors: $x_1 = (0.5, 1.5), x_2 = (1.5, 0.5)$ with $y_1 = y_2 = 1$, and $x_3 = -x_1, x_4 = -x_2$ with $y_3 = y_4 = -1$. The ℓ_2-norm normalized max margin vector is $\widehat{w} = (1, 1)/\sqrt{2}$ with margin equal to $\sqrt{2}$.

Artificial Dataset (Binary Classification)

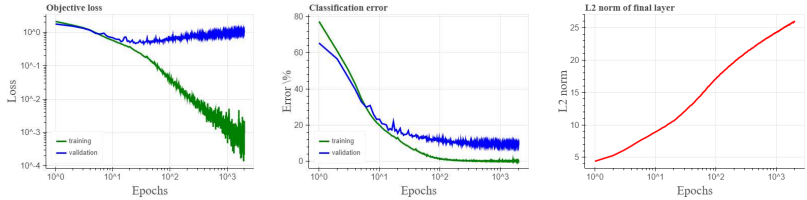


(B): Norm of $w(t)$; (C): Training loss $L\left(w(t)\right)$;

(D & E): the angle and the margin gap.

GDMO: GD with momentum of 0.9.

CIFAR10 (Multiclass Classification)



The sample values can be found in Table 1 the next page.

- From the left to right: loss, classification error and the ℓ_2-norm of the last layer of CNN.
- Color: validation (blue) and training (red).
- Message: after epoch 10, validation loss $\uparrow$ but validation classification error still $\downarrow$.

CIFAR10 (Multiclass Classification)

Epoch	50	100	200	400	2000	4000
L_2 norm	13.6	16.5	19.6	20.3	25.9	27.54
Train loss	0.1	0.03	0.02	0.002	10^{-4}	$3 \cdot 10^{-5}$
Train error	4%	1.2%	0.6%	0.07%	0%	0%
Validation loss	0.52	0.55	0.77	0.77	1.01	1.18
Validation error	12.4%	10.4%	11.1%	9.1%	8.92%	8.9%

Table 1: Sample values from various epochs in the experiment depicted in Fig. 2.

- Validation loss $\uparrow$ but validation classification error still $\downarrow$ after some iterations.

Outline

Multiclass Classification

Consider K-class classification problem and use the generalized logistic loss (i.e., the cross-entropy loss with a softmax output):

$$L\left(\{\mathbf{w}_k\}_{k\in[K]}\right) = -\sum_{n=1}^{N} \log\left(\frac{\exp\left(\mathbf{w}_{y_n}^\top \mathbf{x}_n\right)}{\sum_{k=1}^{K}\exp\left(\mathbf{w}_k^\top \mathbf{x}_n\right)}\right)$$

Theorem 4.

For almost all multiclass datasets (i.e., except for a measure zero) which are linearly separable (i.e., the constrains in Eq. 13 below are feasible), from any starting point $w(0)$ and any small enough stepsize, the iterates of gradient descent will behave as:

$$w_k(t) = \widehat{w}_k \log(t) + \rho_k(t),$$

where the residual $\rho_k(t)$ is bounded and $\widehat{w}_k$ is the solution of the K-class SVM:

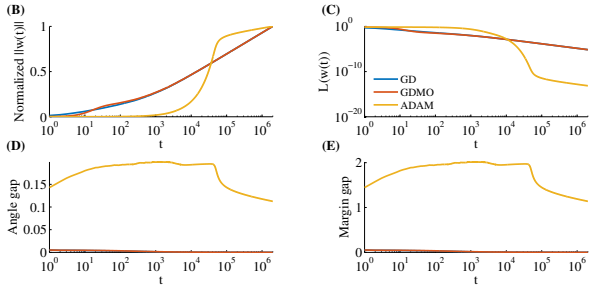
$$\arg\min_{w_1,\cdots,w_k} \sum_{k=1}^{K} \|w_k\|^2 \quad \textit{s.t.} \quad \forall n, \forall k \neq y_n : w_{y_n}^\top x_n \geq w_k^\top x_n + 1. \tag{13}$$

Deep Neural Networks

- A trivial result for DNN is to apply the previous result to the last layer and assume that the previous layers are fixed and linearly separable after some iterations.
- Another reduction is that only a single layer is optimized and the activation units do no switching the sign after some iterations (see Corollary 8 in the original paper).

Adaptive Optimization Methods

- Note that momentum, stochasticity does not change the bias [Nacson et al., 2019].
- However, the implicit bias adaptive methods like Adam depends on the stepsize and initial point, thus is not robust like non-adaptive methods [Wilson et al., 2017, Gunasekar et al., 2018].



Outline

Outline

Proof Sketch For Theorem 1

- For simplicity, let's assume $\ell(u) = \exp(-u)$.
- Asymptotically, $\forall n : w(t)^\top x_n \to \infty$ as $t \to \infty$.
- If $w(t)/\left\|w(t)\right\| \to w_\infty$, then we can write $w(t) = g(t)w_\infty + \rho(t)$ such that $g(t) \to \infty$; moreover, $\forall n : w_\infty^\top x_n \to \infty$ and $\lim_{t\to\infty} \rho(t)/g(t) = 0$.
- The (negative) gradient is:

$$-\nabla L(w) = \sum_{n=1}^{N} \exp\left(-w(t)^\top x_n\right) x_n = \sum_{n=1}^{N} \exp\left(-g(t)w_\infty^\top x_n\right) \exp\left(-\rho(t)^\top x_n\right) x_n. \quad (14)$$

  - As $g(t) \to \infty$, only samples with the least negative exponents will contribute to the gradient.
  - These important samples are actually the smallest margin vectors: $\arg\min_n w_\infty^\top x_n$.

Proof Sketch For Theorem 1

- To prove Theorem 1, we need to show that when $g(t) = \log(t)$, the residual error $\rho(t)$ is relatively small in the sense that the increment in the max margin term is larger than the residual error.
- Therefore, there are two cases: $a)$: $\rho(t)$ is bounded; $b)$: $\rho(t)$ increases slower than $\log(t)$.
- In neither case, we need to show that the increment in the norm of $\rho(t)$ is at most $C_1 t^{-\nu}$ for $C_1 > 0$ and $\nu > 1$, which is a converging series.

A Simple Theorem for Almost Every Dataset

Theorem 5.

For almost every dataset which is linearly separable (see Assumption 1), any β-smooth decreasing loss function (see Assumption 2) with an exponential tail (see Assumption 3), any stepsize $\beta < 2\beta^{-1}\sigma_{\max}^{-2}(X)$ and any starting point $w(0)$, the gradient descent iterates (as in Eq. 7) will behave as:

$$w(t) = \widehat{w}\log(t) + \rho(t),$$

where $\widehat{w}$ is the L_2 max margin vector (the solution to the hard margin SVM):

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^d} \|w\|^2 \quad \textit{s.t.} \quad w^\top x_n \geq 1,$$

and the residual $\rho(t)$ is bounded and so

$$\lim_{t\to\infty} \frac{w(t)}{\|w(t)\|} = \frac{\widehat{w}}{\|\widehat{w}\|}.$$

Proof

Proof of Theorem 5

For any solution $w(t)$, we define

$$r(t) = w(t) - \widehat{w} \log t - \widetilde{w}, \tag{15}$$

where $\widehat{w}$ and $\widetilde{w}$ follow the conditions of Theorem 1 and Theorem 2, respectively. That is, $\widehat{w}$ is the max margin solution and $\widetilde{w}$ is the vector satisfies Eq. 10:

$$\forall n \in \mathcal{S}: \quad \eta \exp\left(-x_n^\top \widetilde{w}\right) = \alpha_n.$$

where $X_{\mathcal{S}} \in \mathbb{R}^{d \times |\mathcal{S}|}$ is the matrix whose columns are the support vectors, a subset $\mathcal{S} \subset \{1, \cdots, N\}$ of the columns of $X = [x_1, \cdots, x_N] \in \mathbb{R}^{d \times N}$.

Proof of Theorem 5

- Based on the KKT condition, for max margin solution $\widehat{w}$:

$$\widehat{w} = \sum_{n=1}^{N} \alpha_n x_n, \quad \forall n \left(\alpha_n \geq 0 \,\&\, \widehat{w}^\top x_n = 1\right) \text{ OR } \left(\alpha_n = 0 \,\&\, \widehat{w}^\top x_n > 1\right). \tag{16}$$

- In Lemma 4 (provided in the auxiliary result), the authors prove that for almost every dataset $\alpha = (\alpha_1, \cdots, \alpha_n)^\top$ is uniquely defined, there are no more than d support vectors and $\alpha_n \neq 0, \forall n \in \mathcal{S}$. Therefore, Eq. 10 is well-defined.
- If the support vectors do not span the data, then the solution $\widetilde{w}$ to Eq. 10 may not be unique. In this case, we can use any such solution in the proof.

Proof of Theorem 5

- We further denote the minimum margin to a non-support vector as:

$$\theta = \min_{n \notin \mathcal{S}} x_n^\top \widehat{w} > 1. \tag{17}$$

- By C_i, ϵ_i, t_i $(i \in \mathbb{N})$, we denote various positive constants which are impendent of t.
- We define $P_1 \in \mathbb{R}^{d \times d}$ as the orthogonal matrix to the subspace spanned by the support vectors, i.e., $P_1 = X_{\mathcal{S}} X_{\mathcal{S}}^+$, where $X_{\mathcal{S}}^+$ is the Moore-Penrose pseudoinverse.
- And $\overline{P_1} = I - P_1$ is the complementary projection (to the left null space of $X_{\mathcal{S}}$).

A Simple Proof for the Exp-loss

- In the following, we first examine the special case that $\ell(u) = \exp(-u)$ and take the continuous time limit of gradient descent: $\eta \to 0$, so
$$\dot{w}(t) = -\nabla L(w(t)).$$
- Note that our goal is to show that $r(t)$:
$$r(t) = w(t) - \log(t)\widehat{w} - \widetilde{w}$$
is bounded so that $\rho(t) = r(t) + \widetilde{w}$ is bounded.
- Note that we have
$$\dot{r}(t) = \dot{w}(t) - \frac{1}{t}\widehat{w} = -\nabla L\left(w(t)\right) - \frac{1}{t}\widehat{w}. \tag{18}$$

A Simple Proof for the Exp-loss

$$
\begin{aligned}
\frac{1}{2}\frac{d}{dt}\left\|r(t)\right\|^2 &= \dot{r}^\top(t) r(t) \\
&= \sum_{n=1}^{N} \exp\left(-x_n^\top w(t)\right) x_n^\top r(t) - \frac{1}{t}\widehat{w}^\top r(t) \\
&= \left[\sum_{n\in\mathcal{S}} \exp\left(-\log(t)\widehat{w}^\top x_n - \widetilde{w}^\top x_n - x_n^\top r(t)\right) - \frac{1}{t}\widehat{w}^\top r(t)\right] \\
&\quad + \left[\sum_{n\notin\mathcal{S}} \exp\left(-\log(t)\widehat{w}^\top x_n - \widetilde{w}^\top x_n - x_n^\top r(t)\right)\right], \qquad (19)
\end{aligned}
$$

where in the last equality we decompose the sum over support vectors $\mathcal{S}$ and non-support vectors and use the relation in Eq. 15.

Proof

A Simple Proof for the Exp-loss

Recall that $\widehat{w}^\top x_n = 1$ for $n \in \mathcal{S}$ and we have defined (in Eq. 10) $\widetilde{w}$ so that (η is missing here?)

$$\sum_{n\in\mathcal{S}} \exp\left(-\widetilde{w}^\top x_n\right) x_n = \widehat{w}.$$

Thus, we can rewrite the first bracketed term in Eq. 19 as:

$$\begin{aligned}
&\frac{1}{t}\sum_{n\in\mathcal{S}} \exp\left(-\widetilde{w}^\top x_n - x_n^\top r(t)\right) - \frac{1}{t}\sum_{n\in\mathcal{S}} \exp\left(-\widetilde{w}^\top x_n\right) x_n^\top r(t) \\
&= \frac{1}{t}\sum_{n\in\mathcal{S}} \left(-\widetilde{w}^\top x_n\right)\left(\exp\left(-x_n^\top r(t)\right) - 1\right) x_n^\top r(t) \\
&\leq 0, \qquad (20)
\end{aligned}$$

where the last inequality is based on $\forall z, z(e^{-z} - 1) \leq 0$.

A Simple Proof for the Exp-loss

For the second bracketed term in Eq. 19, define

$$\theta = \min_{n \notin \mathcal{S}} \widehat{w}^\top x_n > 1.$$

Therefore, the second bracketed term can be upper bounded by

$$\sum_{n \notin \mathcal{S}} \exp\left(-\log(t)\widehat{w}^\top x_n - \widetilde{w}^\top x_n\right) \exp\left(-x_n^\top r(t)\right) x_n^\top r(t) \leq \frac{1}{t^\theta} \sum_{n \notin \mathcal{S}} \exp\left(-\widetilde{w}^\top x_n\right). \qquad (21)$$

where in the last inequality we also use $\forall z, e^{-z} z \leq 1$. Substituting Eq. 20 and 21 to Eq. 19 and integrating, we obtain that $\exists C, C'$ such that

$$\forall t_1, \forall t > t_1 : \quad \|r(t)\|^2 - \|r(t_1)\|^2 \leq C \int_{t_1}^{t_2} \frac{dt}{t^\theta} \leq C' < \infty,$$

Therefore we show that $r(t)$ is bounded, which complete the proof for the special case.

A Complete Proof of Theorem 5

- Now, we consider the general case where step size is not infinitely small and the loss functions are exponentially-tailed.
- We use the following two auxiliary results: Lemma 2 and Lemma 3.

Lemma 2.

Let $L(w)$ be a β-smooth objective function. If $\beta < 2\beta^{-1}$, then from any $w(0)$, with the GD sequence:

$$w(t+1) = w(t) - \eta \nabla L\left(w(t)\right),$$

we have that $\sum_{t=0}^{\infty} \left\|\nabla L\left(w(t)\right)\right\|^2 < \infty$ and therefore $\lim_{t\to\infty} \left\|\nabla L\left(w(t)\right)\right\|^2 = 0$.

A Complete Proof of Theorem 5

Lemma 3.

We have

$$\exists C_1, t_1 : \forall t > t_1 : (r(t+1) - r(t))^\top r(t) \leq C_1 t^{-\min(\theta, 1+1.5\mu_+, 1+0.5\mu_-)}. \tag{22}$$

Additionally, $\forall \epsilon_1 > 0, \exists C_2, t_2 > 0$ *such that* $\forall t > t_2$*, if*

$$\|P_1 r(t)\| \geq \epsilon_1, \tag{23}$$

then the following improved bound holds:

$$(r(t+1) - r(t))^\top r(t) \leq -C_2 t^{-1} < 0. \tag{24}$$

A Complete Proof of Theorem 5

Note that our goal is still to show that $\|r(t)\|$ is bounded and therefore $\rho(t) = r(t) + \widetilde{w}$ is bounded.

$$\|r(t+1)\|^2 = \|r(t+1) - r(t)\|^2 + 2\left(r(t+1) - r(t)\right)^\top r(t) + \|r(t)\|^2. \tag{25}$$

We note the first term can be upper bounded by

$$\begin{aligned}
\|r(t+1) - r(t)\|^2 &= \|w(t+1) - \widehat{w}\log(t+1) - \widetilde{w} - w(t) - \widehat{w}\log(t) + \widetilde{w}\|^2 \\
&= \|-\eta\nabla L\left(w(t)\right) - \widehat{w}\left[\log(t+1) - \log(t)\right]\|^2 \\
&= \eta^2\left\|\nabla L\left(w(t)\right)\right\|^2 + \|\widehat{w}\|^2\log^2(1+t^{-1}) + 2\eta\widehat{w}^\top\nabla L\left(w(t)\right)\log(1+t^{-1}) \\
&\leq \eta^2\left\|\nabla L\left(w(t)\right)\right\|^2 + \|\widehat{w}\|^2 t^{-2}
\end{aligned} \tag{26}$$

where the last inequality we use $\forall x > 0 : x \geq \log(1+x) > 0$ and also

$$\widehat{w}^\top\nabla L\left(w(t)\right) = \sum_{n=1}^{N} \ell'\left(w(t)^\top x_n\right)\widehat{w}^\top x_n \leq 0.$$

A Complete Proof of Theorem 5

From Lemma 2, we know that

$$\|\nabla L\left(w(t)\right)\|^2 = o(1) \quad \text{and} \quad \sum_{t=0}^{\infty} \|\nabla L\left(w(t)\right)\|^2 < \infty.$$

Back to Eq. 26 and recalling that $t^{-\nu}$ power series converges for any $\nu > 1$, we can find C_0 such that

$$\|r(t+1) - r(t)\|^2 = o(1) \quad \text{and} \quad \sum_{t=0}^{\infty} \|r(t+1) - r(t)\|^2 = C_0 < \infty.$$

Note that this equation also implies that $\forall \epsilon_0$:

$$\exists t_0 : \forall t > t_0 : |\|r(t+1)\| - \|r(t)\|| < \epsilon_0. \tag{27}$$

A Complete Proof of Theorem 5

Next, we aim to bound the second term $(r(t+1)-r(t))^\top r(t)$ in Eq. 25. According to the exponentially-tailed property in Lemma 3, we can find t_1, C_1 such that $\forall t > t_1$:

$$(r(t+1)-r(t))^\top r(t) \leq C_1 t^{-\min(\theta, 1+1.5\mu_+, 1+0.5\mu_-)}.$$

Therefore,

$$\begin{aligned}\|r(t)\|^2 - \|r(t_1)\|^2 &= \sum_{u=t_1}^{t-1} \left[\|r(u+1)\|^2 - \|r(u)\|^2\right] \\ &\leq C_0 + 2\sum_{u=t_1}^{t-1} C_1 u^{-\min(\theta, 1+1.5\mu_+, 1+0.5\mu_-)},\end{aligned}$$

which is bounded since $\theta > 1$ and $\mu_-, \mu_+ > 0$. Therefore $\|r(t)\|$ is bounded.

Outline

Generic Solutions of the KKT Coniditions of Max Margin Solution

For the max margin problem:

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^d} \|w\|^2 \quad \text{s.t.} \quad w^\top x_n \geq 1,$$

we can write down the KKT condition:

$$\widehat{w} = \sum_{n=1}^{N} \alpha_n x_n \quad \forall n : \left(\alpha_n \geq 0 \text{ and } \widehat{w}^\top x_n = 1\right) \text{ OR } \left(\alpha_n = 0 \text{ and } \widehat{w}^\top x_n > 1\right). \tag{28}$$

Generic Solutions of the KKT Coniditions of Max Margin Solution

We will show that for support vectors $\alpha_n = 0$ is under measure zero.

Lemma 4.

For almost all datasets there is a unique α which satisfies the KKT conditions (see Eq. 28). Furthermore, in this solution $\alpha_n \neq 0$ if $\widehat{w}^\top x_n = 1$, i.e., x_n is a support vector $(n \in \mathcal{S})$ and there are most d such support vectors.

References I

S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In Proceedings of the 35th International Conference on Machine Learning, pages 1827–1836, 2018.

M. S. Nacson, N. Srebro, and D. Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, pages 3051–3059, 2019.

M. Telgarsky. Margins, shrinkage, and boosting. In Proceedings of the 30th International Conference on Machine Learning, volume 28, pages 307–315, 2013.

A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In Advances in Neural Information Processing Systems 30, pages 4148–4158, 2017.

Proof

References II

T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. The Annals of Statistics, 33(4):1538–1579, 2005.