Benefits of Early Stopping in Gradient Descent for Overparameterized Logistic Regression

Hyojeong An Taesun Yeom

Wu et al., in ICML'25

Overview

1. Preliminaries, Backgrounds, and Setups
2. Part 1: Upper Bounds for Early-Stopped GD
3. Part 2: Early Stopping vs. Asymptotic Regime
4. Part 3: Early Stopping and ℓ_2-Regularization
5. Conclusion

Overview

Motivation

Q. Why do deep neural networks (or at least simplified models) generalize well, even when they are heavily overparameterized?

$\downarrow$

The need for an **easy-to-handle**, **tractable model and setup**.

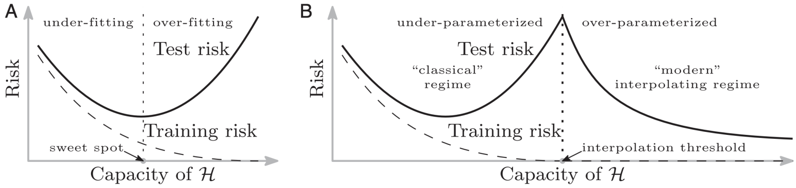

Figure: Deep Double Descent [1]

Back to Basic: Logistic Regression

- Given n data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$

$$\mathbf{x}_i \in \mathbb{R}^d, \quad y_i \in \{+1, 1\}, \quad i \leq n. \tag{1}$$

- Trained with empirical logistic loss

$$\widehat{\mathcal{L}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i \mathbf{x}_i^\top \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})). \tag{2}$$

- Update via a full-batch gradient descent

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \widehat{\mathcal{L}}(\mathbf{w}_t). \tag{3}$$

(Asymptotic) Implicit Bias

- Implicit bias is informally defined as a "characteristic" of the predictor at $t \to \infty$.
  - From regression (i.e., MSE loss) to classification (i.e., exp, logistic loss).

- From [3], if the training data distribution is **linearly separable**, the (normalized) predictor at limit $\tilde{\mathbf{w}}$ becomes ℓ_2 max-margin solution:

$$\tilde{\mathbf{w}} = \arg\max_{\|\mathbf{w}\|=1} \min_i y_i \mathbf{x}_i^\top \mathbf{w} > 0, \quad \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} \to \tilde{\mathbf{w}}. \tag{4}$$

Q. *Does this formula always yield a favorable implicit bias?*
Moreover, is it related to generalization?

Data model

$$\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \quad \Pr(y|\mathbf{x}) = \frac{1}{1 + \exp(-y\mathbf{x}^\top \mathbf{w}^*)}. \tag{5}$$

- $\boldsymbol{\Sigma}$ (covariance matrix): PSD + Bounded trace
- $\mathbf{w}^*$ (true model): $\|\mathbf{w}^*\|_{\boldsymbol{\Sigma}} := {\mathbf{w}^*}^\top \boldsymbol{\Sigma} \mathbf{w}^* < \infty \implies \mathrm{Std}(\mathbf{x}^\top \mathbf{w}^*) < \infty$.
- Labels are generated by $\mathbf{w}^*$ (Eq. 5), with **some inherent noise**.
- During the paper, we will consider the following setup.
    - *Noisy*, $\|\mathbf{w}^*\|_{\boldsymbol{\Sigma}} \lesssim 1$: The true model generates "uncertain" labels (i.e., prediction near 0.5).
    - *Overparameterized*, $\mathrm{rank}(\boldsymbol{\Sigma}) \geq n$: Always exists a hyperplane that **perfectly separates the training set**, i.e., perfect train acc.

Theoretical Picture

Part 1. Early-stopped logistic regression achieves a vanishing risk upper bound.

Part 2. Which outperforms the model in the asymptotic regime.

Part 3. Indeed, there exists a strong connection between early stopping time and ℓ_2-regularization.

Metrics

Let $(\mathbf{x}, y) \in \mathbb{H} \times \{\pm 1\}$. $\mathbb{H}$ is a finite or countably infinite-dimensional Hilbert space.

- **Logistic risk**

$$\mathcal{L}(\mathbf{w}) := \mathbb{E}\ell(y\mathbf{x}^\top \mathbf{w}), \quad \text{where } \ell(t) := \ln(1 + e^{-t}) \tag{6}$$

- Zero-one error

$$\mathcal{E}(\mathbf{w}) := \mathbb{E}\, \mathbf{1}[y\mathbf{x}^\top \mathbf{w} \le 0] = \Pr(y\mathbf{x}^\top \mathbf{w} \le 0) \tag{7}$$

- Calibration error

$$\mathcal{C}(\mathbf{w}) := \mathbb{E}\, |p(\mathbf{w}; \mathbf{x}) - \Pr(y = 1|\mathbf{x})|^2 \tag{8}$$

In this presentation, we will mainly focus on the "logistic risk."

Basic Properties

- **True model is the best model!**
  - The true model $\mathbf{w}^*$ (i.e., Bayes optimal classifier) satisfies

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \quad \text{and} \quad \mathbf{w}^* \in \arg\min_{\mathbf{w}} \mathcal{E}(\mathbf{w}). \tag{9}$$

- **Basic inequality.**
  - The zero-one error is bounded by the calibration error, which is bounded by the logistic risk.

$$\underbrace{\mathcal{E}(\mathbf{w}) - \min \mathcal{E}}_{\text{Zero-one error}} \;\leq\; 2\sqrt{\mathcal{C}(\mathbf{w})} \;\leq\; \sqrt{2 \cdot \underbrace{(\mathcal{L}(\mathbf{w}) - \min \mathcal{L})}_{\text{Logistic risk}}} \tag{10}$$

Overview

1. Preliminaries, Backgrounds, and Setups
2. **Part 1: Upper Bounds for Early-Stopped GD**
3. Part 2: Early Stopping vs. Asymptotic Regime
4. Part 3: Early Stopping and ℓ_2-Regularization
5. Conclusion

Definitions

Based on the Properties, minimizing logistic risk leads to the following properties:
An estimator $\widehat{\mathbf{w}}$ is called

- **Calibrated** if $\mathcal{C}(\widehat{\mathbf{w}}) \to 0$.
  $\Rightarrow$ The model's predicts the true one.
- **Consistent** if $\mathcal{E}(\widehat{\mathbf{w}}) - \min \mathcal{E} \to 0$.
  $\Rightarrow$ The model achieves Bayes optimal zero-one error.

Theorem

Early-Stopped GD successfully minimizes the excess logistic risk, thereby achieving both calibration and consistency.

A bias-dominating risk bound

Theorem

Let k be an arbitrary index. Suppose that the stepsize η for GD satisfies $\eta \leq 1/(C_0(1 + tr(\mathbf{\Sigma}) + \lambda_1 \ln(1/\delta)/n))$ where $C_0 > 1$ is a universal constant. Then with probability at least $1 - \delta$, there exists a stopping time t such that

$$\widehat{\mathcal{L}}(\mathbf{w}_t) \leq \widehat{\mathcal{L}}(\mathbf{w}^*_{0:k}) \leq \widehat{\mathcal{L}}(\mathbf{w}_{t-1}). \tag{11}$$

Moreover, for GD with this stopping time, we have

$$\mathcal{L}(\mathbf{w}_t) - \min \mathcal{L} \leq \sqrt{\max\left\{1, tr(\mathbf{\Sigma})\|\mathbf{w}^*_{0:k}\|^2\right\} \frac{\ln^2(n/\delta)}{n}} + \frac{1}{2}\|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{\Sigma}}, \tag{12}$$

where C is a constant.

Setup

Let $(\lambda_i, \mathbf{u}_i)$ be the eigenvalue, eigenvector pair of the covariance matrix $\mathbf{\Sigma}$. Let $\pi(i)$ be resorted indexes such that $\lambda_{\pi(i)}(\mathbf{u}_{\pi(i)}^\top \mathbf{w}^*)^2$ is non-increaing fuction. Define

$$\mathbf{w}_{0:k}^* := \sum_{i=1}^{k} \mathbf{u}_{\pi(i)} \mathbf{u}_{\pi(i)}^\top \mathbf{w}^*$$

$$\mathbf{w}_{k:\infty}^* := \sum_{i=k+1}^{\infty} \mathbf{u}_{\pi(i)} \mathbf{u}_{\pi(i)}^\top \mathbf{w}^*$$

- We pick k which satisfies $\|\mathbf{w}_{0:k}^*\| = o(\sqrt{n})$.
  Then the **risk bound** implies that **logistic risk** vanishes:

$$\mathcal{L}(\mathbf{w}_t) - \min \mathcal{L} = o(1) \quad \text{as } n \text{ increases}$$

Proof Sketch (1)

To bound the risk $\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^*)$, we decompose it into four terms using the empirical risk $\widehat{\mathcal{L}}$ and the model $\mathbf{w}^*_{0:k}$.

$$\begin{aligned}\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^*) = &\underbrace{[\mathcal{L}(\mathbf{w}_t) - \widehat{\mathcal{L}}(\mathbf{w}_t)]}_{\text{(1: Lemma B.3)}} + \underbrace{[\widehat{\mathcal{L}}(\mathbf{w}_t) - \widehat{\mathcal{L}}(\mathbf{w}^*_{0:k})]}_{\text{(2)}} \\ &+ \underbrace{[\widehat{\mathcal{L}}(\mathbf{w}^*_{0:k}) - \mathcal{L}(\mathbf{w}^*_{0:k})]}_{\text{(3: Lemma B.3)}} + \underbrace{[\mathcal{L}(\mathbf{w}^*_{0:k}) - \mathcal{L}(\mathbf{w}^*)]}_{\text{(4: Lemma B.2)}}\end{aligned}$$

- We define the stopping time t such that $\widehat{\mathcal{L}}(\mathbf{w}_t) \leq \widehat{\mathcal{L}}(\mathbf{w}^*_{0:k})$, then term (2) ≤ 0.
- Before calculating generalization error, we must ensure the early-stopped parameter $\mathbf{w}_t$ does not explode.

Boundedness

Lemma B.1

Let $\beta := C_0(1 + \operatorname{tr}(\boldsymbol{\Sigma}) + \lambda_1 \ln(1/\delta)/n)$, where $C_0 > 1$ is a sufficiently large constant. Assume that $\eta \le 1/\beta$ and t is such that $\widehat{\mathcal{L}}(\mathbf{w}^*_{0:k}) \le \widehat{\mathcal{L}}(\mathbf{w}_{t-1})$. Then with probability at least $1-\delta$, we have $\|\mathbf{w}_t - \mathbf{w}^*_{0:k}\| \le 1 + \|\mathbf{w}^*_{0:k}\|$.

Note that

$$\|\mathbf{w}_t - \mathbf{w}^*_{0:k}\| \le \|\mathbf{w}_t - \mathbf{w}_{t-1}\| + \|\mathbf{w}_{t-1} - \mathbf{w}^*_{0:k}\|$$

- $\widehat{\mathcal{L}}$ is $\sqrt{\beta}$-Lipschitz
- Let $\widehat{\mathcal{L}}(\cdot)$ be convex and β-smooth. Then for every $\mathbf{u}$, we have:

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta t} + \widehat{\mathcal{L}}(\mathbf{w}_t) \le \widehat{\mathcal{L}}(\mathbf{u}) + \frac{\|\mathbf{u}\|^2}{2\eta t} \tag{13}$$

Proof Sketch (2)

We need to show that $\widehat{\mathcal{L}}$ is β-smooth and $\sqrt{\beta}$-Lipschitz for $\beta > 1$
Note that

$$\|\nabla \widehat{\mathcal{L}}(\mathbf{w})\| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2}, \quad \|\nabla^2 \widehat{\mathcal{L}}(\mathbf{w})\| \leq \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2.$$

By Bernstein's inequality, we have the following with probability at least $1-\delta$:

$$\begin{aligned}
\sum_{i=1}^{n}\|\mathbf{x}_i\|^2 = \sum_{i=1}^{n}\sum_{j}\lambda_j z_{ij}^2 &\leq n\mathrm{tr}(\boldsymbol{\Sigma}) + C_1\left(\sqrt{n\sum_{j}\lambda_j^2\ln(1/\delta)} + \lambda_1\ln(1/\delta)\right) \\
&\leq C_0(n\mathrm{tr}(\boldsymbol{\Sigma}) + \lambda_1\ln(1/\delta)) \leq \beta,
\end{aligned}$$

where $C_0, C_1 > 1$ are constants.

Generalization Error

Lemma B.3

Let $C_1 > 1$ be a sufficiently large constant. Then with probability at least $1 - \delta$,

$$\sup_{\|\mathbf{w}\| \le W} |\mathcal{L}(\mathbf{w}) - \widehat{\mathcal{L}}(\mathbf{w})| \le C_1 W \sqrt{\frac{(1 + \mathrm{tr}(\boldsymbol{\Sigma})) \ln(n/\delta) \ln(1/\delta)}{n}}.$$

By applying Rademacher Complexity Bounds [2], we have

$$\begin{aligned}\sup_{\|\mathbf{w}\| \le W} (L(\mathbf{w}) - \widehat{\mathcal{L}}(\mathbf{w})) &\le \frac{W}{n} + 2XW\sqrt{\frac{1}{n}} + XW\sqrt{\frac{\ln(1/(3\delta))}{2n}} \\ &\le C_1 W \sqrt{\frac{(1 + \mathrm{tr}(\boldsymbol{\Sigma})) \ln(n/\delta) \ln(1/\delta)}{n}},\end{aligned}$$

where $C_1 > 1$ is a constant.

Approximation Error

Lemma B.2

Let $\mathbf{w}^* \in \arg\min \mathcal{L}(\mathbf{w})$, then for every $\mathbf{w}$, we have

$$\mathcal{L}(\mathbf{w}) \leq \mathcal{L}(\mathbf{w}^*) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{\Sigma}}^2. \tag{14}$$

Notice that

$$\nabla^2 \mathcal{L}(\mathbf{w}) = \mathbb{E}\ell''(y\mathbf{x}^\top \mathbf{w})\mathbf{x}\mathbf{x}^\top = \mathbb{E}\frac{\mathbf{x}\mathbf{x}^\top}{(1+\exp(\mathbf{x}^\top \mathbf{w}))(1+\exp(-\mathbf{x}^\top \mathbf{w}))} \preceq \mathbb{E}\mathbf{x}\mathbf{x}^\top = \mathbf{\Sigma}. \tag{15}$$

Moreover, we have $\nabla \mathcal{L}(\mathbf{w}^*) = 0$. Then by the midpoint theorem, there exists a $\mathbf{v}$ such that

$$\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}^*) = \langle \nabla \mathcal{L}(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \nabla^2 \mathcal{L}(\mathbf{v})(\mathbf{w} - \mathbf{w}^*) \leq \frac{1}{2}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{\Sigma}}^2. \tag{16}$$

Overview

1. Preliminaries, Backgrounds, and Setups
2. Part 1: Upper Bounds for Early-Stopped GD
3. **Part 2: Early Stopping vs. Asymptotic Regime**
4. Part 3: Early Stopping and ℓ_2-Regularization
5. Conclusion

Asymptotic Behavior of the Logistic Risk

Theorem (Logistic risk at limit)

Suppose that Assumption 1 holds. Let $\tilde{\mathbf{w}}$ *be a unit vector such that* $\|\tilde{\mathbf{w}}\|_{\boldsymbol{\Sigma}} > 0$ *and let* $(\mathbf{w}_t)_{t\geq 0}$ *be a sequence of vectors such that*

$$\|\mathbf{w}_t\| \to \infty, \ \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} \to \tilde{\mathbf{w}}. \tag{17}$$

Then we have

$$\mathcal{L}(\mathbf{w}_t) = \infty. \tag{18}$$

If data are "noisy"...

- Logistic risk ($\mathcal{L}$) at limit: Diverges!

Early stopping is better than asymptotic GD!

Proof Sketch (1)

Fix a small constant $\gamma > 0$. Define an *special* event

$$\mathcal{F} := \{\mathbf{x} : |\mathbf{x}^\top \mathbf{w}^*| \leq 10\|\mathbf{w}^*\|_{\mathbf{\Sigma}},\ \|\mathbf{x}\| \leq 10\sqrt{\mathrm{tr}(\mathbf{\Sigma})},\ |\mathbf{x}^\top \tilde{\mathbf{w}}| \geq \gamma\}. \tag{19}$$

Let t_0 be such that $\|\mathbf{w}_t/\|\mathbf{w}_t\| - \tilde{\mathbf{w}}\| \leq \gamma/\left(20\sqrt{\mathrm{tr}(\mathbf{\Sigma})}\right)$, for every $t \geq t_0$. Next, we have

$$\frac{|\mathbf{x}^\top \mathbf{w}_t|}{\|\mathbf{w}_t\|} \geq |\mathbf{x}^\top \tilde{\mathbf{w}}| - \left|\mathbf{x}^\top \left(\frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} - \tilde{\mathbf{w}}\right)\right| \tag{20}$$

$$\geq |\mathbf{x}^\top \tilde{\mathbf{w}}| - \|\mathbf{x}\| \left\|\frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} - \tilde{\mathbf{w}}\right\| \geq \frac{\gamma}{2}. \tag{21}$$

Proof Sketch (2)

Then for every $\mathbf{x} \in \mathcal{F}$ and $t \geq t_0$, we have the population risk

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_t) &= \mathbb{E}_{\mathbf{x}} \mathbb{E}_y \ln(1 + \exp(-y\mathbf{x}^\top \mathbf{w}_t)) && (22) \\
&\geq \mathbb{E} \sum_{y \in \{-1,+1\}} \frac{\ln(1 + \exp(-y\mathbf{x}^\top \mathbf{w}_t))}{1 + \exp(-y\mathbf{x}^\top \mathbf{w}^*)} \mathbf{1}[\mathbf{x} \in \mathcal{F}] && (23) \\
&\geq \mathbb{E} \frac{\ln(1 + \exp(|\mathbf{x}^\top \mathbf{w}_t|))}{1 + \exp(|\mathbf{x}^\top \mathbf{w}^*|)} \mathbf{1}[\mathbf{x} \in \mathcal{F}] && (24) \\
&\geq \mathbb{E} \frac{\|\mathbf{w}_t\| \gamma / 2}{1 + \exp(10 \|\mathbf{w}^*\|_{\mathbf{\Sigma}})} \Pr(\mathcal{F}) \to \infty. && (25)
\end{aligned}$$

Overview

Early stopping vs. ℓ_2-regularization

Theorem (Early stopping vs. ℓ_2-reg.)

Let $\widehat{\mathcal{L}}(\cdot)$ be convex and β-smooth. Consider $(\mathbf{w}_t)_{t\geq 0}$ given by GD iteration with $\eta \leq 1/\beta$ and $(\mathbf{u}_\lambda)_{\lambda>0} := \arg\min_{\mathbf{u}} \widehat{\mathcal{L}}(\mathbf{u}) + (\lambda/2)\|\mathbf{u}\|^2$. Set $\lambda = 1/(\eta t)$. Then we have ,

$$\|\mathbf{w}_t - \mathbf{u}_\lambda\| \leq \frac{1}{\sqrt{2}}\|\mathbf{w}_t\|. \tag{26}$$

As a direct consequence, we also have

$$\cos(\mathbf{w}_t, \mathbf{u}_\lambda) \geq \frac{1}{\sqrt{2}}, \qquad \frac{\sqrt{2}}{\sqrt{2}+1}\|\mathbf{u}_\lambda\| \leq \|\mathbf{w}_t\| \leq \frac{\sqrt{2}}{\sqrt{2}-1}\|\mathbf{u}_\lambda\|. \tag{27}$$

- Under certain settings, the trajectories of ℓ_2-reg. and early stopping do not differ significantly in terms of angle and norm.
- $\lambda = 1/(\eta t)$: **Short training (i.e., small t) $\iff$ Strong reg. (i.e., large λ)**.

Proof Sketch (1)

For the first-order stationary point of the ℓ_2-reg. ERM

$$\nabla \widehat{\mathcal{L}}(\mathbf{u}_\lambda) + \lambda \mathbf{u}_\lambda = \nabla \widehat{\mathcal{L}}(\mathbf{u}_\lambda) + \frac{1}{\eta t}\mathbf{u}_\lambda. \tag{28}$$

With the results of the early stopping + convexity, we have,

$$\begin{aligned} \frac{1}{2}\|\mathbf{w}_t - \mathbf{u}_t\|^2 - \frac{1}{2}\|\mathbf{u}_t\|^2 &\leq \eta t(\widehat{\mathcal{L}}(\mathbf{u}_\lambda) - \widehat{\mathcal{L}}(\mathbf{w}_t)) && (29) \\ &\leq \eta t \langle \nabla \mathcal{L}(\mathbf{u}_\lambda), \mathbf{u}_\lambda - \mathbf{w}_t \rangle = -\langle \mathbf{u}_\lambda, \mathbf{u}_\lambda - \mathbf{w}_t \rangle. && (30) \end{aligned}$$

Rearranging the terms, we get

$$\frac{1}{2}\|\mathbf{w}_t - \mathbf{u}_\lambda\|^2 \leq \langle \mathbf{u}_\lambda, \mathbf{w}_t \rangle - \frac{1}{2}\|\mathbf{u}_\lambda\|^2 \iff \|\mathbf{w}_t - \mathbf{u}_\lambda\|^2 \leq \frac{1}{2}\|\mathbf{w}_t\|^2. \tag{31}$$

Proof Sketch (2)

1. Angle

From $2\langle \mathbf{u}_\lambda, \mathbf{w}_t \rangle \geq \frac{1}{2}\|\mathbf{w}_t\|^2 + \|\mathbf{u}_\lambda\|^2$, we obtain

$$\cos(\mathbf{u}_\lambda, \mathbf{w}_t) = \frac{\langle \mathbf{u}_\lambda, \mathbf{w}_t \rangle}{\|\mathbf{u}_\lambda\|\|\mathbf{w}_t\|} \geq \frac{\frac{1}{2}\left(\frac{1}{2}\|\mathbf{w}_t\|^2 + \|\mathbf{u}_\lambda\|^2\right)}{\|\mathbf{u}_\lambda\|\|\mathbf{w}_t\|} \geq \frac{1}{\sqrt{2}}. \tag{32}$$

2. Norm

We have

$$\frac{1}{\sqrt{2}}\|\mathbf{w}_t\| \geq \|\mathbf{w}_t - \mathbf{u}_\lambda\| \geq \begin{cases} \|\mathbf{w}_t\| - \|\mathbf{u}_\lambda\| \\ \|\mathbf{u}_\lambda\| - \|\mathbf{w}_t\|, \end{cases} \tag{33}$$

which implies

$$\frac{\sqrt{2}}{\sqrt{2}+1}\|\mathbf{u}_\lambda\| \leq \|\mathbf{w}_t\| \leq \frac{\sqrt{2}}{\sqrt{2}-1}\|\mathbf{u}_\lambda\|. \tag{34}$$

Blessing of Dimensionality?

Q. What if we care about the ℓ_2 distance between the predictors?

- In high dimension, the "suport vector condition" typically holds, which is defined as

$$\mathrm{rank}\{x_i : i \in \mathcal{S}_+\} = \mathrm{rank}\{x_i : i \in [n]\}. \tag{35}$$

$=$ the rank of the support vector should be same as the rank of the data matrix.

- Under this condition, we have,

$$\boxed{\|\mathbf{w}_t - \mathbf{u}_{\lambda(t)}\| \to 0,} \quad \text{while } \|\mathbf{w}_t\|,\ \|\mathbf{u}_{\lambda(t)}\| \to \infty, \quad \text{as } t \to \infty. \tag{36}$$

- But with low-dimensional vectors, there exists some exceptions, i.e., difference between the norm diverges.

Overview

Conclusion & Discussion

So far, this work investigates:

- Early stopping in overparameterized logistic regression improves logistic risk, zero-one error, and calibration error,
- Early stopping is better than asymptotic GD,
- Strong connection between early stopping and ℓ_2-regularization.

However, several challenges remain to be addressed in future work:

- The results depends on the oracle-chosen stopping time.
- This work focuses on linear models, not deep nets.
  - Networks with more than two layers exhibit distinct and more complex phenomena (e.g., as the lazy-rich dynamic [4]).

Q & A

References

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal.
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
Proceedings of the National Academy of Sciences, 2019.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari.
On the complexity of linear prediction: Risk bounds, margin bounds, and regularization.
Advances in neural information processing systems, 2008.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro.
The implicit bias of gradient descent on separable data.
Journal of Machine Learning Research, 2018.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro.
Kernel and rich regimes in overparametrized models.
In *Conference on Learning Theory*, 2020.