A review of uniform convergence of the landscape of empirical risk functions

Torey Hilbert

December 7, 2021

1 Introduction

We first describe the generic empirical risk minimization problem [MBM17]. A sample $\{z_1, \ldots, z_n\} \subseteq \mathbb{R}^d$ is drawn from a distribution $\mathbb{P}_Z$. Furthermore, with $\theta \in \Theta_{n,p} \subseteq \mathbb{R}^p$ and $z \in \mathbb{R}^d$, we have a loss function $\ell : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$. The loss function gives rise to two other functions; the empirical risk,

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; z_i),$$

and the population risk

$$R(\theta) = E_{\mathbb{P}_Z}[\ell(\theta, Z)].$$

To try to estimate θ that minimizes $\mathbb{R}(\theta)$, we construct the estimator

$$\hat{\theta}_n = \underset{\theta \in \Theta_{n,p}}{argmin} \hat{R}_n(\theta).$$

An existing result tells us that this is a reasonable approach.

Theorem 1 ([Gee00])**.** *If Θ is compact, $\ell(\theta; Z)$ is continuous in θ, and*

$$E_Z \left[\underset{\theta \in \Theta}{sup}\ \ell(\theta; Z) \right] < \infty,$$

then the empirical risk converges uniformly to the population risk; that is,

$$\underset{\theta \in \Theta}{sup} |\hat{R}_n(\theta) - R(\theta)| \xrightarrow{P} 0,$$

i.e. for any $\epsilon, \delta > 0$ there is some $N \in \mathbb{N}$ such that for $n \geq N$, with probability at least $1 - \delta$,

$$\underset{\theta \in \Theta}{sup} |\hat{R}_n(\theta) - R(\theta)| < \epsilon.$$

Intuitively, this result tells us that if we minimize the empirical risk for a large enough sample, then we should expect that the minimum value of the empirical risk should be close to the minimum value of the population risk. However, this does not strictly preclude the possibility that the critical points of $\hat{R}_n$ and R are significantly different; namely $\nabla\hat{R}_n(\theta)$ could vary a lot in regions where $\nabla R(\theta)$ is relatively constant. This so-called "landscape" of the empirical risk function is the study of these two papers [MBM17; ZF17].

2 Contributions of [MBM17]

In this paper, uniform convergence results for both $\nabla\hat{R}_n(\theta)$ and $\nabla^2\hat{R}_n(\theta)$ are constructed for specific situations, under moderately strict regularity conditions on Z, and for $\theta \in \Theta_{n,p} = B^p(r)$ for some fixed $r > 0$. In particular, they make three major assumptions on Z:

- $\nabla\ell(\theta; Z)$ is τ-sub-Gaussian.
- For any λ with $||\lambda||_2 \leq 1$, $\nabla^2\ell(\theta; Z)$ is τ^2-sub-Exponential.
- There is some constant c_h such that

$$E_Z\left[\sup_{\theta_1\neq\theta_2\in B^p(r)} \frac{||\nabla^2\ell(\theta_1; Z) - \nabla^2\ell(\theta_2; Z)||_{op}}{||\theta_1 - \theta_2||_2}\right] \leq J_* \leq \tau^3 p^{c_h}.$$

Under these assumptions, they prove the following theorem:

Theorem 2 ([MBM17])**.** *For any $\delta > 0$, there is a constant C_0 such that if we define $C = C_0\,(c_h \vee log(r\tau/\delta) \vee 1)$, then for any $n \geq Cplogp$, we have*

$$P\left(\sup_{\theta\in B^p(r)} ||\nabla\hat{R}_n(\theta) - \nabla R(\theta)||_2 \leq \tau\sqrt{\frac{Cplogn}{n}}\right) \geq 1 - \delta$$

and

$$P\left(\sup_{\theta\in B^p(r)} ||\nabla^2\hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \leq \tau^2\sqrt{\frac{Cplogn}{n}}\right) \geq 1 - \delta.$$

These results tell us that the gradient of the empirical risk uniformly converges to the gradient of the population risk as we increase the sample size, and the Hessian of the empirical risk uniformly converges to the Hessian of the population risk as we increase the sample size. Furthermore, this convergence happens in $O\left(\sqrt{\frac{logn}{n}}\right)$.

The general idea of the proof is to use an ϵ-cover $\{\theta_1, \ldots, \theta_N\}$ on the parameter space $\Theta_{n,p}$ to obtain something of the form

$$\begin{aligned}||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 \leq & ||\frac{1}{n}\sum_i [\nabla \ell(\theta; Z_i) - \nabla \ell(\theta_{j(\theta)}; Z_i)]||_2 \\ & + ||\frac{1}{n}\sum_i \nabla \ell(\theta_{j(\theta)}; Z_i) - E_Z[\nabla \ell(\theta_{j(\theta)}; Z)]||_2 \\ & + ||E_Z[\nabla \ell(\theta_{j(\theta)}; Z] - E_Z[\nabla \ell(\theta; Z)]||_2,\end{aligned}$$

so that

$$P\left(\sup_{\theta \in B^p(r)} ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 \geq t\right) \leq P(A_t) + P(B_t) + P(C_t),$$

where A_t is the event

$$\sup_{\theta \in B^p(r)} ||\frac{1}{n}\sum_i [\nabla \ell(\theta; Z_i) - \nabla \ell(\theta_{j(\theta)}; Z_i||_2 \geq t/3,$$

B_t is the event

$$\sup_{j=1,\ldots,N} ||\frac{1}{n}\sum_i \nabla \ell(\theta_{j(\theta)}; Z_i) - E_Z[\nabla \ell(\theta_{j(\theta)}; Z)||_2 \geq t/3,$$

and C_t is the event

$$\sup_{\theta \in B^p(r)} ||E_Z[\nabla \ell(\theta_{j(\theta)}; Z] - E_Z[\nabla \ell(\theta; Z)]||_2 \geq t/3.$$

Here $j(\theta)$ denotes any point θ_j in the ϵ-cover that covers θ, i.e. $\theta \in B^p(\theta_{j(\theta)}, \epsilon)$. B_t is bounded using a union bound and a technical lemma about coverings of $B^p(1)$, taking advantage of the fact that we're only taking the supremum over a finite set. In particular, bounding B_t uses the sub-Gaussian assumption on the gradient. C_t is a deterministic event that's bounded using the operator norm bound assumptions on the Hessian of the empirical risk. Lastly, A_t is also bounded using the operator norm bound on the Hessian on the empirical risk. The proof for the uniform convergence of the Hessian of the empirical risk is similar in style albeit using different bounds.

The paper also proves a strong structural theorem that vaguely states that, under the same assumptions as the previous theorem, with probability at least $1-\delta$, for large n the critical points of $\hat{R}_n(\theta)$ and $R(\theta)$ on $\Theta_{n,p}$ are in one-to-one correspondence with each other. For even larger n depending on the operator norm of the third derivative of $R(\theta)$, $\sup_{\theta \in B^p(r)} ||\nabla^3 R(\theta)||_{op}$, we have that each critical point of $\hat{R}_n(\theta)$ is close to the associated critical point of $R(\theta)$; that is, with probability at least $1-\delta$,

$$||\hat{\theta}_n^{(j)} - \theta^{(j)}||_2 \leq \frac{2\tau}{\eta}\sqrt{\frac{Cplogn}{n}}.$$

The paper demonstrates the use of these two theorems by giving theoretical guarantees for three practice problems. I will focus on their first example - logistic regression with a non-convex loss function in the high dimensional setting. In particular, let $\theta_0 \in \mathbb{R}^d$, and let $\sigma : \mathbb{R} \to [0,1]$. Then suppose that $Y \in \{0,1\}$, $X \in \mathbb{R}^d$ and $P(Y = 1|X = x) = \sigma(\theta_0^T x)$. Lastly, suppose that we are given n pairs $(y_1, x_1), \ldots, (y_n, x_n)$ from $P_{X,Y}$. We want to minimize

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sigma(\theta^T x_i)\right)^2,$$

over the set $\Theta_{n,d} = B^d(r)$.

They specifically make the following three assumptions, as well:

- The first three derivatives of σ are bounded, and that $\sigma\prime(z) > 0$ for all $z \in \mathbb{R}$.
- X has mean zero and is τ-sub-Gaussian.
- $E_X[XX^T]$ is positive definite, i.e. X spans R^d.

Then they get the following result on the effectiveness of gradient descent on this empirical risk function:

Theorem 3 ([MBM17]). *Under the given assumptions, if $||\theta_0||_2 \leq r/3$, then there are constants C_1, C_2, and h_{max} independent of n and d such that if $n \geq C_1 dlogd$, then with probability at least $1 - \delta$:*

- *The empirical risk function $\hat{R}_n(\theta)$ has a unique local minimum in $B^d(r)$.*
- *Gradient descent with fixed step size $h \leq h_{max}$ converges exponentially to the global minimum for any initialization $\theta_s \in B^d(\theta_0, 2r/3)$.*
- $||\hat{\theta}_n - \theta_0||_2 \leq C_2\sqrt{(dlogn)/n}$.

These three results provide good justification for the general framework of the empirical risk minimization problem, and for the usage of gradient descent even when we have non-convex loss functions. One future direction is to specialize these results to situations that arise more commonly in practice; [ZF17] pursues this by formulating and proving theorems of similar nature to those given above for specifically deep linear and deep non-linear neural networks. Beyond that, one could study how the local connectivity found in CNNs affects the landscape and convergence rates, given that many applications take advantage of convolutional layers. Lastly in this direction, one could study how frequently the sub-Gaussian/sub-exponential assumptions on the gradient and Hessian of the loss functions arise in practice.

Another interesting future direction would be to study the robustness of the landscape of the empirical risk when a few of the sample points do not follow the same distribution as $\mathbb{P}_Z$. On a similar note, studying the expected value $E_Z[\sup_{\theta \in B^p(r)} ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2]$ instead of bounding the probability

of it exceeding a fixed value could be interesting as well. Lastly, for practical purposes, methods of estimating a reasonable value for $r > 0$ for a given dataset could be helpful.

3 My contribution

The results of the experiments found in section 5.1 of [MBM17] suggest studying the behavior of $\nabla \hat{R}_n(\theta)$ as $||\theta||_2 \to \infty$. In particular, we consider the following problem:

Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be a sigmoid function. Let $\theta_0 \in R^d$. Then suppose that $Y \in \{0, 1\}$ and $P(Y = 1|X = x) = \sigma(\theta_0^T x)$. Lastly, suppose that we are given n pairs $(y_1, x_1), \ldots, (y_n, x_n)$. Then we have the following result.

Theorem 4. *If $n \geq 2$, $d \geq 2$, and there exists x_i in the sample such that no nonzero scalar multiples of x_i are in the sample, then there is some direction $\lambda \in \mathbb{R}^d$, $||\lambda||_2 = 1$, such that for some $t_0 \in \mathbb{R}$, $\lambda^T \nabla \hat{R}_n(t\lambda) < 0$ for all $t > t_0$.*

Intuitively, this result tells us that for any real dataset, there will always be some direction where, far enough away from the origin, gradient descent will push the parameter away from the origin and out towards ∞. This justifies the assumption of [MBM17] to only attempt inference on compact sets $\Theta_{n,p}$.

Proof. Define $z_i = x_i$ if $y_i = 0$, and $z_i = -x_i$ if $y_i = 1$. Then notice that $1 - \sigma(x) = \sigma(-x)$ for all $x \in \mathbb{R}$, so

$$\begin{aligned}
(y_i - \sigma(\theta^T x_i))^2 &= \begin{cases} (-\sigma(\theta^T x_i))^2 & \text{if } y_i = 0, \\ (\sigma(-\theta^T x_i))^2 & \text{if } y_i = 1 \end{cases} \\
&= \begin{cases} \sigma(\theta^T x_i)^2 & \text{if } y_i = 0, \\ \sigma(-\theta^T x_i)^2 & \text{if } y_i = 1 \end{cases} \\
&= \begin{cases} \sigma(\theta^T z_i)^2 & \text{if } y_i = 0, \\ \sigma(\theta^T z_i)^2 & \text{if } y_i = 1 \end{cases} \\
&= \sigma(\theta^T z_i)^2.
\end{aligned}$$

For ease of notation, define $\rho(w) = \sigma(w)^2$, so that

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sigma(\theta^T x_i))^2 = \frac{1}{n} \sum_{i=1}^{n} \rho(\theta^T z_i).$$

By direct computation, $\rho'(w) = \frac{2e^{-w}}{(1+e^{-w})^3}$, so

$$\rho'(w) \geq \frac{e^{2w}}{4} \text{ if } w \leq 0$$

and

$$\rho'(w) \leq 2e^{-w} \text{ if } w > 0.$$

For any direction $\lambda \in \mathbb{R}^p$, $||\lambda||_2 = 1$, define $w_i = \lambda^T z_i$ to be the component of z_i in the direction of λ. We consider what happens to $\nabla \hat{R}_n(t\lambda)$ as $t \to \infty$. Define

$$\begin{aligned} I &= \{i \in \{1, \dots, n\} \mid w_i > 0\}, \\ J &= \{i \in \{1, \dots, n\} \mid w_i < 0\}, \\ K &= \{i \in \{1, \dots, n\} \mid w_i = 0\}. \end{aligned}$$

Then for $t > 0$,

$$\begin{aligned} \lambda^T \nabla \hat{R}_n(t\lambda) &= \frac{1}{n} \sum_{i=1}^{n} \rho'(t\lambda^T z_i) \lambda^T z_i \\ &= \frac{1}{n} \sum_{i=1}^{n} \rho'(tw_i) w_i \\ &= \frac{1}{n} \left(\sum_{i \in I} \rho'(tw_i) w_i + \sum_{i \in J} \rho'(tw_i) w_i + \sum_{i \in K} \rho'(tw_i) w_i \right) \\ &\leq \frac{1}{n} \left(\sum_{i \in I} 2e^{-tw_i} w_i + \sum_{i \in J} \frac{e^{2tw_i}}{4} w_i \right) \end{aligned}$$

Let $w_+ = \min_{i \in I} w_i$ and $w_- = \max_{j \in J} w_j$.

$$\begin{aligned} \lim_{t \to \infty} \left| \frac{\sum_{i \in I} 2e^{-tw_i} w_i}{\sum_{i \in J} \frac{e^{2tw_i}}{4} w_i} \right| &= \lim_{t \to \infty} -\frac{8e^{-tw_+} w_+}{e^{2tw_-} w_-} \frac{\sum_{i \in I} e^{-t(w_i - w_+)} \frac{w_i}{w_+}}{\sum_{i \in J} e^{2t(w_i - w_-)} \frac{w_i}{w_-}} \\ &= -\frac{8w_+}{w_-} \lim_{t \to \infty} e^{-t(2w_- + w_+)} \\ &= \begin{cases} 0 & \text{if } 2w_- + w_+ > 0, \\ \infty & \text{if } 2w_- + w_+ < 0. \end{cases} \end{aligned}$$

In particular, if $|w_+| > 2|w_-|$, then there is some $t_0 > 0$ such that for all $t > t_0$,

$$\left| \frac{\sum_{i \in I} 2e^{-tw_i} w_i}{\sum_{i \in J} \frac{e^{2tw_i}}{4} w_i} \right| < 1,$$

so that $|\sum_{i \in I} 2e^{-tw_i} w_i| < |\sum_{i \in J} \frac{e^{2tw_i}}{4} w_i|$, so that

$$\lambda^T \nabla \hat{R}_n(t\lambda) \leq \frac{1}{n} \left(\sum_{i \in I} 2e^{-tw_i} w_i + \sum_{i \in J} \frac{e^{2tw_i}}{4} w_i \right) < 0.$$

All that remains is to find such a λ. First notice that if any z_i are exactly zero, we can ignore them since then $w_i = \lambda^T z_i = 0$ for all $\lambda \in \mathbb{R}$. Hence WLOG assume that all z_i are nonzero. By hypothesis, there is some x_j, with

$j \in \{1, \ldots, n\}$, so that $z_i \neq cz_j$ for any $c \in \mathbb{R}$ and $i \neq j$. Pick $\lambda_0 \in \mathbb{R}$, $||\lambda_0||_2 = 1$ in the hyperplane orthogonal to z_j and such that λ_0 is not orthogonal to any other points z_i in the sample (recall that $d \geq 2$, so that this is possible).

Then construct λ by rotating λ_0 away from z_j by a small enough amount such that $2|\lambda^T z_j| < |\lambda^T z_i|$ for all $i \neq j$ (this is possible because $|\lambda_0^T z_i| > c > 0$ for some constant $c > 0$ for all $i \neq j$, since the z_i are not orthogonal to λ_0). Defining the $w_i = \lambda^T z_i$ as above, we find that $w_j = \lambda^T z_j < 0$, and in particular $w_- = w_j$ and $2|w_-| < |w_i|$ for all $i \neq j$, so $2|w_-| < |w_+|$. Hence there is some $t_0 > 0$ so that for all $t > t_0$, $\lambda^T \nabla \hat{R}_n(t\lambda) < 0$, as desired. □

References

[Gee00] Sara van de Geer. *Empirical Process Theory and Applications*. Cambridge University Press, 2000.

[MBM17] Song Mei, Yu Bai, and Andrea Montanari. *The Landscape of Empirical Risk for Non-convex Losses*. 2017. arXiv: `1607.06534 [stat.ML]`.

[ZF17] Pan Zhou and Jiashi Feng. *The Landscape of Deep Learning Algorithms*. 2017. arXiv: `1705.07038 [stat.ML]`.