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

Torey Hilbert

November 2021

Problem Background

We first describe the 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 $\theta \in \Theta_{n,p} \subseteq \mathbb{R}^p$, 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 estimate θ that minimizes $R(\theta)$, we construct the estimator

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

Existing Results on Uniform Convergence

Theorem ([Gee00])

If Θ is compact, $\ell(\theta; Z)$ is continuous in θ, and

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

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

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

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

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

Existing Results on Uniform Convergence

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 object of study in these two papers [MBM17; ZF17].

Contributions of [MBM17]

In [MBM17], 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

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

Uniform Convergence of Gradient

Under aforementioned assumptions, they prove the following theorem:

Theorem ([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

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

Proof Sketch (abbreviated from [MBM17])

Pick an ϵ-cover $\{\theta_1, \ldots, \theta_N\}$ of the parameter space $\Theta_{n,p}$. Let $j(\theta)$ denote any point θ_j in the ϵ-cover that covers θ, i.e. $\theta \in B^p(\theta_{j(\theta)}, \epsilon)$.

We use triangle inequality to break up the desired event into the three events A_t, B_t, and C_t as follows:

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

and hence

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

Convergence of Critical Points

They also proves a structural theorem, vaguely stated as follows:

Theorem ([MBM17])

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 $1-\delta$,

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

Binary Classification Example [MBM17]

[MBM17] considers three main example applications of their results. The one that I will discuss is on logistic regression with a nonconvex loss function in a high dimensional setting.

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

Binary Classification Theorem

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.

Theorem ([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 min in $B^d(r)$.*
- *For any initialization $\theta_s \in B^d(\theta_0, 2r/3)$, gradient descent with fixed step size $h \leq h_{max}$ converges exponentially to the global min.*
- $||\hat{\theta}_n - \theta_0||_2 \leq C_2\sqrt{(dlogn)/n}$.

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 $\theta \to \infty$. In particular, we consider the following problem:
Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be a sigmoid function, and suppose that we are given n pairs of points $(y_1, x_1), \ldots, (y_n, x_n)$. Define the empirical risk function

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

Then we have the following result.

Theorem

If $n \geq 2$, $d \geq 2$, and there exists x_i in the sample such that no 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 $t > t_0$.

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].