1 CLT for Logistic Regression

Definition 1. If X_n is a sequence of random variables with cdfs F_n, X is a random variable with cdf F, and $F_n(x) \to F(x)$ for all points x where F is continuous, then X_n *converges in law* to X; in symbols, $X_n \xrightarrow{L} X$.

Theorem 1. *Let $X_1, X_2, ..., X_n$ be iidrv with density $f_\theta(x)$ satisfying the following conditions:*

1. *The distributions P_θ are distinct. That is, $P_{\theta_1} = P_{\theta_2}$ implies that $\theta_1 = \theta_2$.*
2. *The parameter space $\theta \in \Omega$ is open.*
3. *The density $f_\theta(x)$ is continuous in x.*
4. *The set $A = \{x : f_\theta(x) > 0\}$ is independent of θ.*
5. *For all $x \in A$, $f_\theta(x)$ is three times differentiable with respect to θ, and the third derivative is continuous. The corresponding derivatives of the integral $\int f_\theta(x)dx$ can be obtained by differentiating under the integral sign.*
6. *If θ_0 denotes the true value of θ, there exists a positive number $c(\theta_0)$ and a function $M_{\theta_0}(x)$ such that*

$$\left| \frac{\partial^3}{\partial \theta^3} \log f_\theta(x) \right| \leq M_{\theta_0}(x) \textit{ for all } x \in A, |\theta - \theta_0| < c(\theta_0) \tag{1}$$

and

$$E_{\theta_0}[M_{\theta_0}(X)] < \infty \tag{2}$$

Then any consistent sequence $\hat{\theta}_n = \hat{\theta}(X_1, ..., X_n)$ of roots of the likelihood equation satisfies

$$\hat{\theta} \xrightarrow{L} \theta_0 + \frac{1}{\sqrt{n}} \mathcal{N}\left(0, I^{-1}(\theta_0)\right) \tag{3}$$

where $I(\theta_0)$ is the Fisher information.

Logistic regression with the L_2 loss satisfies conditions 1-6 above. The L_1 loss does not satisfy the conditions above because it is not everywhere differentiable. To work around this limitation, define the function

$$R_\alpha(\theta) = \sum_i \sqrt{\alpha \theta_i^2 + 1} - 1 \tag{4}$$

where θ_i is the ith component of θ. This function is three times differentiable, and it converges to the L_1 norm as $\alpha \to \infty$. We can now perform the analysis using the R_α norm as an arbitrarily close approximation to the L_1 norm.

2

Lemma 1 ([2])**.** *Let $A \in \mathbb{R}^{n\times n}$ be a matrix, and let $\Sigma = A^{\mathsf{T}}A$. Let $x = (x_1, ..., x_d)$ be an isotropic multivariate Gaussian random vector with zero mean. For all $t > 0$,*

$$\Pr\left[|Ax|^2 > \operatorname{tr}\Sigma + 2\sqrt{\operatorname{tr}(\Sigma^2)t} + 2|\Sigma|t\right] \leq e^{-t} \tag{5}$$

In the special case where A is the identity, this simplifies to

$$\Pr\left[|x|^2 > d + 2\sqrt{dt} + 2t\right] \leq e^{-t} \tag{6}$$

Lemma 2. *Let $\mathbf{w}_1, ..., \mathbf{w}_m$ be a sequence of $m > 2$ random d-dimensional vectors sampled independently from the isotropic normal distribution. Define $H = \{\sum_{i=1}^m \alpha_i \mathbf{w}_i : \sum_{i=1}^m \alpha_i = 1\}$ to be the smallest hyperplane containing $\mathbf{w}_1, ..., \mathbf{w}_m$, and $h = |\pi_H \mathbf{0}|$ to be the minimum distance from H to the origin. Then,*

$$\Pr\left[h^2 < d - m + 2 + 2\sqrt{(d-m+2)t} + 2t\right] \geq 1 - e^{-t} \tag{7}$$

Proof. Fix $\mathbf{w}_1, ..., \mathbf{w}_{m-1}$. Let G be the smallest hyperplane containing $\mathbf{w}_1, ..., \mathbf{w}_{m-1}$, U be the corresponding vector subspace, and $U^* = \operatorname{span}\{\pi_G \mathbf{0}\}$. Define the point $\mathbf{x} = \pi_U \mathbf{w}_m - \pi_G \pi_U \mathbf{w}_m + \pi_G \mathbf{0}$. The vector $\pi_U \mathbf{w}_m - \pi_G \pi_U \mathbf{w}_m$ is in U^*, so $\mathbf{x}$ is also in U^*. Since U^* is a line whose direction is independent of $\mathbf{x}$, $\mathbf{x}$ is distributed according to a one dimensional standard normal distribution. Also by construction, we have that $\mathbf{x} + \pi_{U^\perp}\mathbf{w}_m \in H$. This implies that $h < |\mathbf{x} + \pi_{U^\perp}\mathbf{w}_m|$. This vector has dimension $d - m + 2$ and an isotropic normal distribution. Applying Lemma 1 gives the result.

□

Lemma 3 ([1])**.** *Let $X_1, ..., X_d$ be d independent Gaussian $\mathcal{N}(0,1)$ random variables, and let $Y = \frac{1}{|X|}(X_1, ..., X_d)$. Let the vector $Z \in \mathbb{R}^k$ be the projection of Y onto its first k coordinates, and let $L = |Z|^2$. Clearly, $\mathbb{E}[L] = k/d$. If $k < d$, then*

1. *If $\beta < 1$, then*

$$\Pr\left[L \leq \frac{\beta k}{d}\right] \leq \beta^{k/2}\left(1 + \frac{(1-\beta)k}{d-k}\right)^{(d-k)/2} \leq \exp\left(\frac{k}{2}(1 - \beta + \ln\beta)\right) \tag{8}$$

2. *If $\beta > 1$, then*

$$\Pr\left[L \geq \frac{\beta k}{d}\right] \leq \beta^{k/2}\left(1 + \frac{(1-\beta)k}{d-k}\right)^{(d-k)/2} \leq \exp\left(\frac{k}{2}(1 - \beta + \ln\beta)\right) \tag{9}$$

Lemma 4. *Let $\mathbf{w}^*$ be an arbitrary d dimensional vector, and $\mathbf{w}_1, ..., \mathbf{w}_m$ be a sequence of $m > 2$ random d-dimensional vectors sampled independently from the isotropic normal distribution. Define $H = \{\sum_{i=1}^m \alpha_i \mathbf{w}_i : \sum_{i=1}^m \alpha_i = 1\}$ to be the smallest hyperplane containing $\mathbf{w}_1, ..., \mathbf{w}_m$. Then for all $\beta > 1$ and $t > 0$,*

$$\Pr\left[|\mathbf{w}^* - \pi_H \mathbf{w}^*|^2 \le |\mathbf{w}^*| \left(1 - \left(\frac{\beta m}{d}\right)\right) + d - m + 2 + 2\sqrt{(d-m+2)t} + 2t\right] \\ \ge \left(1 - e^{-t}\right) \operatorname{erf}\left(\frac{\beta m}{d}\right) \left(1 - \exp\left(\frac{m}{2}(1 - \beta + \ln \beta)\right)\right)^2 \tag{10}$$

Proof. Fix $\mathbf{w}_1, ..., \mathbf{w}_{m-1}$. Let G be the smallest hyperplane containing $\mathbf{w}_1, ..., \mathbf{w}_{m-1}$, and U be the corresponding vector subspace. We have that

$$|\mathbf{w}^* - \pi_H \mathbf{w}^*| \le |\mathbf{w}^* - \pi_H \pi_U \mathbf{w}^*| \qquad \text{by definition of } \pi_H \tag{11}$$

$$\le |\mathbf{w}^* - \pi_U \mathbf{w}^*| + |\pi_U \mathbf{w}^* - \pi_H \pi_U \mathbf{w}^*| \qquad \text{by triangle ineq.} \tag{12}$$

We will bound each of these terms separately.

We begin with the first term by noting that the vectors $(\mathbf{w}^* - \pi_U \mathbf{w}^*)$ and $\pi_U \mathbf{w}^*$ are orthogonal. This lets us use the Pythagorean theorem to conclude that

$$|\mathbf{w}^* - \pi_U \mathbf{w}^*| = \sqrt{|\mathbf{w}^*|^2 - |\pi_U \mathbf{w}^*|^2} \tag{13}$$

The vector $\pi_U \mathbf{w}^*$ is a fixed vector projected onto a random subspace, which has the same distribution as a random vector projected onto a fixed subspace. Therefore, we can apply Lemma 3 to get

$$\Pr\left[|\pi_U \mathbf{w}^*| \le |\mathbf{w}^*| \left(\frac{\beta m}{d}\right)\right] \ge 1 - \exp\left(\frac{m}{2}(1 - \beta + \ln \beta)\right) \tag{14}$$

Combining Equations 13 and 14 gives

$$\Pr\left[|\mathbf{w}^* - \pi_U \mathbf{w}^*| \le |\mathbf{w}^*| \left(1 - \left(\frac{\beta m}{d}\right)\right)\right] \ge 1 - \exp\left(\frac{m}{2}(1 - \beta + \ln \beta)\right) \tag{15}$$

Now for the second term. Define the line $U^* = \{\alpha \pi_U \mathbf{w}^* + (1-\alpha)\pi_G \pi_U \mathbf{w}^*\}$, and the point $\mathbf{x} = \pi_U \mathbf{w}_m - \pi_G \pi_U \mathbf{w}_m + \pi_G \mathbf{w}^*$. By construction, we have that $\mathbf{x} \in U^*$ and $\mathbf{x} + \pi_{U^\perp} \mathbf{w}_m \in H$.

$$|\pi_U \mathbf{w}^* - \pi_H \pi_U \mathbf{w}^*| \le |\pi_U \mathbf{w}^* - \pi_H \mathbf{x}| \qquad \text{by definition of } \pi_H \tag{16}$$

$$|\pi_U \mathbf{w}^* - \pi_H \pi_U \mathbf{w}^*|^2 = |\pi_U \mathbf{w}^* - \mathbf{x}|^2 + |\mathbf{x} - \pi_H \mathbf{x}|^2 \qquad \text{by Pythagorean theorem} \tag{17}$$

$$\le |\pi_U \mathbf{w}^* - \mathbf{x}|^2 + |\mathbf{x} - (\mathbf{x} + \pi_{U^\perp} \mathbf{w}_m)|^2 \qquad \text{by definition of } \pi_H \tag{18}$$

$$= |\pi_U \mathbf{w}^* - \mathbf{x}|^2 + |\pi_{U^\perp} \mathbf{w}_m|^2 \tag{19}$$

The right vector above is normally distributed, but the left vector is not. Our strategy will be to bound the left vector in probability by a normally distributed vector, then apply Lemma 1 to the result. In particular, $|\pi_{U^*}\mathbf{0} - \mathbf{x}|$ has a standard normal distribution, and $|\pi_{U^*}\mathbf{0} - \mathbf{x}| \geq |\mathbf{x} - \pi_U \mathbf{w}^*|$ whenever $|\pi_{U^*}\mathbf{0} - \mathbf{x}| \geq |\pi_{U*}\mathbf{0} - \pi_U \mathbf{w}^*|$. By the definition of a normal distribution, we have

$$\Pr\left[|\pi_{U^*}\mathbf{0} - \mathbf{x}| \geq \frac{\beta m}{d}\right] \geq \operatorname{erf}\left(\frac{\beta m}{d}\right) \tag{20}$$

Furthermore, we have that $|\pi_{U^*}\mathbf{0} - \pi_U \mathbf{w}^*| \leq |\pi_U \mathbf{w}^*|$, which is upper bounded in probability by Equation 14. Combining Equations 14, 19, and 20 gives:

$$\begin{aligned}\Pr&\left[|\pi_U \mathbf{w}^* - \pi_H \pi_U \mathbf{w}^*|^2 \leq |\pi_U \mathbf{w}^* - \mathbf{x}|^2 + |\pi_{U^\perp} \mathbf{w}_m|^2\right] \\ &\geq \operatorname{erf}\left(\frac{\beta m}{d}\right)\left(1 - \exp\left(\frac{m}{2}(1 - \beta + \ln \beta)\right)\right)\end{aligned} \tag{21}$$

Now combining Equation 21 above with Lemma 1 gives our final bound on the right hand term:

$$\begin{aligned}\Pr&\left[|\pi_U \mathbf{w}^* - \pi_H \pi_U \mathbf{w}^*|^2 \leq d - m + 2 + 2\sqrt{(d-m+2)t} + 2t\right] \\ &\geq (1 - e^{-t}) \operatorname{erf}\left(\frac{\beta m}{d}\right)\left(1 - \exp\left(\frac{m}{2}(1 - \beta + \ln \beta)\right)\right)\end{aligned} \tag{22}$$

$\square$

3 Parallelization

In everything that follows, we assume the likelihood function f satisfies the CLT conditions and already incorporates the regularization penalty.

Let there be m machines we are parallelizing over. All previous work assumes that the data on each machine follows the same distribution. In this analysis, we will relax that assumption. For each machine i, let D_i be the distribution of data assigned to that machine.

$$X_i \sim D_i^{n_i}; X = (X_1, X_2, ..., X_m)$$
$$X_i' \sim D_i^{n_i}; X' = (X_1', X_2', ..., X_m')$$

3.1 Baseline approach

Define $\mathbf{w}$ to be the parameters from training on the entire dataset. That is,

$$\mathbf{w} = \arg\max_{\mathbf{w}} \sum_{x \in X} f(x; \mathbf{w}) \tag{23}$$

By the CLT, we get the convergence rate

$$\mathbf{w} \xrightarrow{L} \mathbf{w}^* + \frac{1}{\sqrt{nm}} \mathcal{N}\left(0, I^{-1}(\mathbf{w}^*)\right) \tag{24}$$

3.2 Averaging

The averaging parallel algorithm has relatively poor asymptotic convergence, and when the D_i distributions are different can converge to an arbitrarily bad value.

For each machine i, train $\mathbf{w}_i$ on the machine's local dataset only. That is,

$$\mathbf{w}_i = \arg\max_{\mathbf{w}} \sum_{x \in X_i} f(x; \mathbf{w}) \tag{25}$$

According to the CLT, we get the convergence rate

$$\mathbf{w}_i \xrightarrow{L} \mathbf{w}_i^* + \frac{1}{\sqrt{n}} \mathcal{N}\left(0, I^{-1}(\mathbf{w}_i^*)\right) \tag{26}$$

Merge the results according to the formula

$$\bar{\mathbf{w}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{w}_i \tag{27}$$

Combining equations 26 and 27 yields

$$\bar{\mathbf{w}} \xrightarrow{L} \bar{\mathbf{w}}^* + \frac{1}{\sqrt{n}} \mathcal{N}\left(0, \frac{1}{m} \sum_{i=1}^{m} I^{-1}(\mathbf{w}_i^*)\right); \bar{\mathbf{w}}^* = \frac{1}{m} \sum_{i=1}^{m} \mathbf{w}_i^* \tag{28}$$

This method is not consistent because $\bar{\mathbf{w}}^*$ need not equal $\mathbf{w}^*$.

3.3 Nested optimizations

Define the projection matrix

$$W = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_m) \tag{29}$$

then this method merges by solving the optimization problem over data points projected onto W. That is,

$$\tilde{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{x \in X'} f(Wx; \mathbf{w}) \tag{30}$$

The vector $\tilde{\mathbf{w}}$ only has dimension m. The final solution is given by projecting back into the original space: $W^T \tilde{\mathbf{w}}$.

Note that the summation is over the data points in X', not in X. This is important to ensure that the projected data points Wx are independent, which is required for the CLT. So by the CLT, we get the convergence rate

$$\tilde{\mathbf{w}} \xrightarrow{L} \tilde{\mathbf{w}}^* + \frac{1}{\sqrt{nm}} \mathcal{N}\left(0, I^{-1}(\tilde{\mathbf{w}}^*)\right) \tag{31}$$

We are interested in

$$W^T \tilde{\mathbf{w}}^* - \mathbf{w}^* = \tag{32}$$

References

[1] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[2] Daniel Hsu, Sham M Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52):1–6, 2012.