

# GLM Hessian Notes

Damian Pavlyshyn

November 6, 2018

## 1 Definitions and set-up

We will consider a glm to be a collection of samples  $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathbf{R}^p \times \mathbf{R}$ , where

$$y_i|x_i \sim P_{x_i^T \beta}$$

for  $\{P_\eta|\eta \in H\}$  a 1-parameter exponential family indexed by a natural parameter  $\eta$  with densities

$$p_\eta(y) = \exp\{\eta y - \psi(\eta)\}p_0(y).$$

Conditional on  $X$ , the matrix whose rows are the  $x_i$ s,  $y$  thus has density

$$f_\beta(y) = \exp\left\{\beta^T(X^T y) - \sum_{i=1}^n \psi(x_i^T \beta)\right\}f_0(y).$$

The likelihood and corresponding derivatives are then given by

$$\begin{aligned}\ell(\beta) &= \beta^T(X^T y) - \sum_{i=1}^n \psi(x_i^T \beta) + \log f_0(y), \\ \nabla \ell(\beta) &= X^T y - \sum_{i=1}^n \psi'(x_i^T \beta)x_i, \\ \nabla^2 \ell(\beta) &= -\sum_{i=1}^n x_i \psi''(x_i^T \beta)x_i^T \\ &= -X^T D_{X\beta} X,\end{aligned}$$

where  $D_{X\beta}$  is the diagonal matrix with  $i$ th entry  $\psi''(x_i^T \beta)$ , which is the conditional variance of  $y_i$ .

Notice in particular that the Hessian of the log-likelihood has no dependence on  $y$  — a feature unique to GLMs.

## 2 The Hessian at the MLE

We are interested in the spectrum of  $\nabla^2 \ell(\hat{\beta})$ . Notice from the previous remark that this depends on  $y$  only through  $\hat{\beta}$ . In particular, if  $\hat{\beta}$  is close to the true  $\beta_0$ , we can expect that  $\nabla^2 \ell(\hat{\beta})$  is close to  $\nabla^2 \ell(\beta_0)$ , and so doesn't depend on  $y$ . In this case, the Hessian would be unaffected by using  $y$  to learn  $\beta$ .

To show this convergence, we will study the distance between eigenvalue distributions of matrices of the form  $X^T D X$ . In particular, for any Lipschitz function  $f$ , we have that

$$\begin{aligned} \left| \int f(\lambda) d\mu_u(\lambda) - \int f(\lambda) d\mu_v(\lambda) \right| &\leq \frac{1}{n} \sum_{i=1}^n |f(\lambda_i) - f(\nu_i)| \\ &\leq \frac{1}{n} \sum_{i=1}^n |\lambda_i - \nu_i| \\ &\leq \frac{1}{n} \|X^T (D_u - D_v) X\|_1 \\ &\leq \frac{1}{n} \|D_u - D_v\|_1 \|X\|_\infty^2 \\ &= \|u - v\|_1 \frac{1}{n} \|X\|_\infty^2. \end{aligned}$$

If  $X$  is an  $n \times p$  matrix of iid normals,  $n \rightarrow \infty$  with  $p$  fixed, we have that  $\|X\|_\infty^2 \sim n$  and  $\|\hat{\beta} - \beta_0\| \rightarrow 0$ . In this case, we have that

$$\mu_{\hat{\beta}} \rightarrow \mu_{\beta_0}$$

almost surely in Wasserstein distance (and hence, for example, weakly).

## 3 Logistic regression

Consider the logistic setting where  $y_i$  takes values  $\{\pm 1\}$ . In this case, we have that

$$\psi''(\eta) = \frac{1}{\cosh^2 \eta},$$

and therefore

$$\nabla^2 \ell(\beta) = - \sum_{i=1}^n \frac{x_i x_i^T}{\cosh^2 x_i^T \beta}.$$

For simplicity, consider the case where  $x_{ij}$  are iid normals and  $\beta \in \text{Uniform}(S^{p-1})$  is constant. Let  $P = I - \beta \beta^T$  be the projection onto the subspace orthogonal to  $\beta$ . We can then write

$$\nabla^2 \ell(\beta) = - \sum_{i=1}^n \frac{(P x_i) x_i^T}{\cosh^2 x_i^T \beta} - \beta \sum_{i=1}^n \frac{(\beta^T x_i) x_i^T}{\cosh^2 x_i^T \beta}.$$

Asymptotically, the second

## 4 Broad strokes argument for semicircle law for small covariance matrices

First, notice that for a matrix  $A$  and scalar  $\lambda$ , we have that

$$\begin{aligned} s_{\lambda A}(z) &= \frac{1}{d} \operatorname{tr}(\lambda A - zI)^{-1} \\ &= \lambda^{-1} \frac{1}{d} \operatorname{tr}(A - (z/\lambda)I)^{-1} \\ &= \lambda^{-1} s_A(z\lambda^{-1}). \end{aligned}$$

From this it follows that

$$z_{\lambda A}(s) = \lambda z_A(\lambda s), \quad R_{\lambda A}(s) = \lambda R_A(\lambda s).$$

Now, we have that, asymptotically,  $\|x_i\|^2 \sim p$ . Hence, we write that

$$\begin{aligned} s_{x_i x_i^T/p}(z) &= \frac{1}{p} \operatorname{tr} \left( \frac{1}{p} x_i x_i^T - zI \right)^{-1} \\ &= \frac{1}{p} \left[ \frac{1}{\|x_i\|^2/p - z} - \frac{p-1}{z} \right] \\ &\approx \frac{1}{p} \left[ \frac{1}{1-z} - \frac{p-1}{z} \right]. \end{aligned}$$

For large  $p$ , it follow that then

$$\begin{aligned} z_{x_i x_i^T/p}(s) &\approx -\frac{1}{s} + \frac{1}{p(1+s)}, \\ R_{x_i x_i^T/p}(s) &\approx \frac{1}{p(1-s)}. \end{aligned}$$

Assuming the corresponding convolution is indeed asymptotically free, we have that

$$\begin{aligned} R_{\sum_{i=1}^n x_i x_i^T/p}(s) &\approx \frac{n}{p(1-s)}, \\ R_{\frac{1}{2\sqrt{np}} \sum_{i=1}^n x_i x_i^T}(s) &= \frac{1}{2} \sqrt{\frac{p}{n}} R_{\sum_{i=1}^n x_i x_i^T/p} \left( \frac{1}{2} \sqrt{\frac{p}{n}} s \right) \\ &\approx \sqrt{\frac{n}{p}} \frac{1}{2 - s\sqrt{n/p}}. \end{aligned}$$

Finally, since  $R_I(z) = 1$ , we have that

$$\begin{aligned} R_{\frac{1}{2\sqrt{np}} (\sum_{i=1}^n x_i x_i^T - nI)}(s) &\approx \sqrt{\frac{n}{p}} \frac{1}{2 - s\sqrt{p/n}} - \frac{1}{2} \sqrt{\frac{n}{p}} \\ &= \sqrt{\frac{n}{p}} \frac{s\sqrt{p/n}}{2(2 - s\sqrt{p/n})} \end{aligned}$$

$$\begin{aligned}
 &= \frac{s}{2(2 - s\sqrt{p/n})} \\
 &\approx \frac{s}{4}.
 \end{aligned}$$

Indeed, this is the  $R$ -transform of the semicircle law on  $[-1, 1]$ . Notice that it seems that the  $x_i$  having internal dependence causes no problems, except possibly rendering the sum asymptotically free as long as  $\|x_i\|^2 \sim p$ .