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

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

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

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

Conditional on X, the matrix whose rows are the x_is, y thus has density

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

The likelihood and corresponding derivatives are then given by

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

where $D_{X\beta}$ is the diagonal matrix with ith entry $\psi''(x_i^{\mathrm{T}}\beta)$, 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 β_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 β.

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

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

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

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

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

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

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

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

Asymptotically, the second

4 Broad strokes argument for semicircle law for small covariance matrices

First, notice that for a matrix A and scalar λ, 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

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

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

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

For large p, it follow that then

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

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

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

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

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

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

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

To generalise, suppose that instead, $\|x_i\|^2 \sim pW_i$ with $W_i > 0$ and $\mathbf{E}W_i = \mu$. We then have that

$$
R_{x_i x_i^{\mathrm{T}}/p}(s) \approx \frac{w_i}{p(1 - sw_i)},
$$

which renders, for large n,

$$
\begin{aligned}
R_{\sum_{i=1}^n x_i x_i^{\mathrm{T}}/p}(s) &= \sum_{i=1}^n \frac{w_i}{p(1 - sw_i)} \\
&\approx \frac{n}{p}\mathbf{E}\Big[\frac{W}{1 - sW}\Big], \\
R_{\frac{1}{2\sqrt{np}}\sum_{i=1}^n x_i x_i^{\mathrm{T}}}(s) &\approx \sqrt{\frac{n}{p}}\mathbf{E}\Big[\frac{W}{2 - sW\sqrt{p/n}}\Big], \\
R_{\frac{1}{2\sqrt{np}}(\sum_{i=1}^n x_i x_i^{\mathrm{T}} - n\mu I)}(s) &\approx \sqrt{\frac{n}{p}}\bigg(\mathbf{E}\Big[\frac{W}{2 - sW\sqrt{n/p}}\Big] - \frac{\mu}{2}\bigg) \\
&= \sqrt{\frac{n}{p}}\mathbf{E}\Big[\frac{sW^2\sqrt{p/n}}{2(2 - sW\sqrt{n/p})}\Big] \\
&\approx \frac{s}{4}\mathbf{E}W^2 \\
&= \frac{(s\sqrt{\mathbf{E}W^2})}{4}\sqrt{\mathbf{E}W^2}.
\end{aligned}
$$

Thus, we have that the ESD of

$$
\frac{1}{\sqrt{\mathbf{E}W^2}}\frac{1}{2\sqrt{np}}\bigg(\sum_{i=1}^n x_i x_i^{\mathrm{T}} - n\mu I\bigg)
$$

converges to the semicircle law.

We see this result supported empirically in fig. 1 which shows the abovse scaling applied to the matrix XDX^{T}, where X are $p \times n$ iid standard Gaussians, D is a diagonal matrix with diagonal entries $1/\cosh^2 W_i$, for standard normal W_i and $n = 30000, p = 300$.

Applying this result to the case of logistic regression, we expect to see a single eigenvalue at $-n\mathbf{E}(W^2/\cosh^2 W)$ and the bulk conforming to a semicircle law supported on $n\mu \pm 2\sigma\sqrt{np}$. With the same parameters as in the previous simulation, we see in fig. 2 that this is largely accurate.

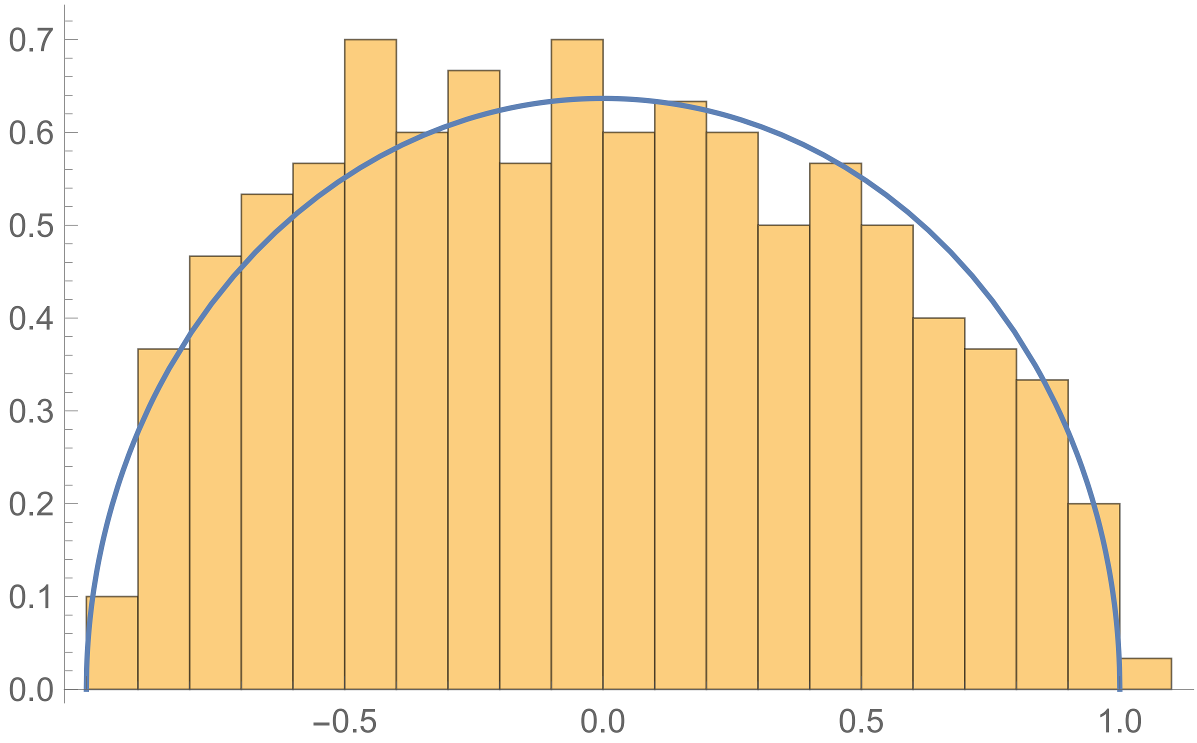

Figure 1: Eigenvalues of a scaled and centred sample covariance matrix.

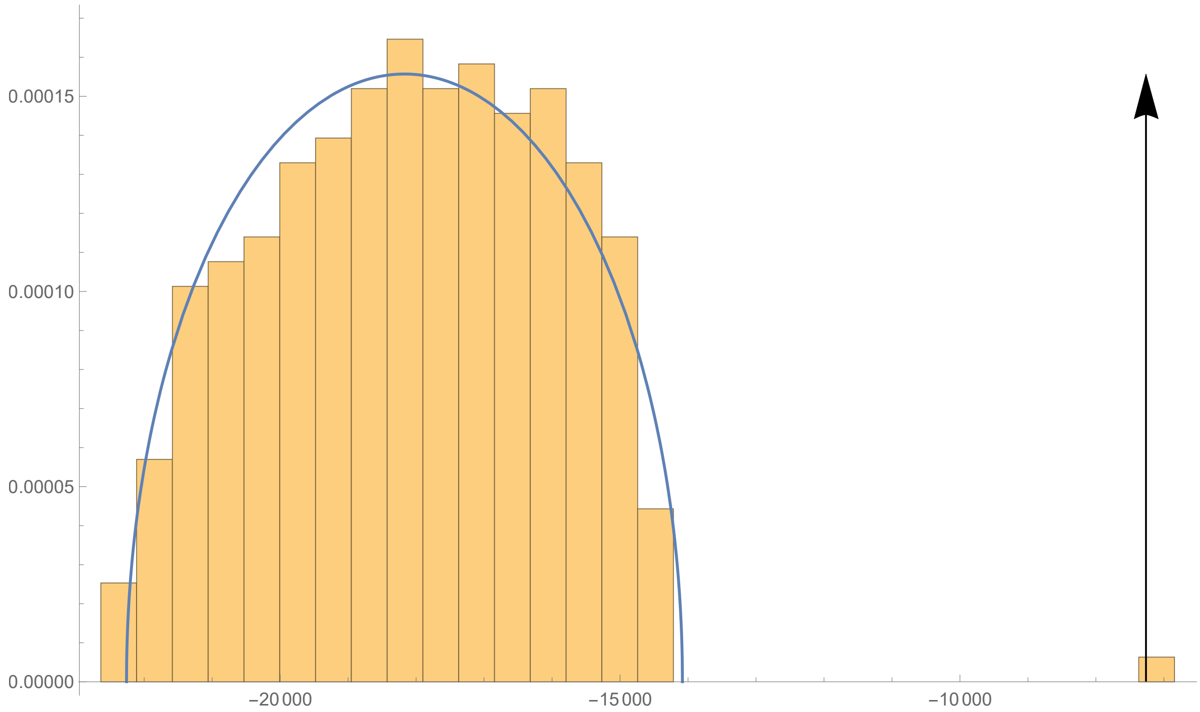

Figure 2: Eigenvalues of the Hessian of logistic regression and theoretical predictions.