Something something Higgs boson

...

January 2, 2021

Abstract

...

Appendices

A AMS

Is the AMS metric convex?

$$\text{AMS} = \sqrt{2(s+b)\log(1+\frac{s}{b}) - s}$$

In the documentation, the implication is that in general, we expect $b \geqq s$ giving the approximate AMS as AMS $\sim \frac{\text{s}}{\sqrt{\text{b}}}(1+\mathcal{O}(\frac{\text{s}}{\text{b}}))$.

If we directly check the derivatives of the approximate AMS in the large b regime:

$$\begin{aligned}
\partial_s^2 \frac{s}{\sqrt{b}} &= 0 \\
\partial_b^2 \frac{s}{\sqrt{b}} &= \frac{3}{4}\frac{s}{b^{5/2}} > 0 \\
\partial_{bs}^2 &= -\frac{1}{2b^{3/2}} < 0
\end{aligned}$$

which implies non-convexity (and non-concavity).

To check more carefully though, calculate the terms of the Hessian of the original and then apply the approximation:

$$\begin{aligned}
\partial_s \text{AMS} &= \log(1+\frac{s}{b})\text{AMS}^{-1} \\
\partial_b \text{AMS} &= (\log(1+\frac{s}{b}) + sb)\text{AMS}^{-1} \\
\partial_s^2 \text{AMS} &= \frac{1}{b+s}\text{AMS}^{-1} - \log(1+\frac{\text{s}}{\text{b}})^2\text{AMS}^{-3} \\
\partial_b^2 \text{AMS} &= (\frac{sb}{s+b} + s)\text{AMS}^{-1} - (\log(1+\frac{\text{s}}{\text{b}}) + \text{sb})^2\text{AMS}^{-3} \\
\partial_{sb}^2 &= (\frac{1}{s+b} + b)\text{AMS}^{-1} - \log(1+\frac{\text{s}}{\text{b}})(\log(1+\frac{\text{s}}{\text{b}}) + \text{sb})\text{AMS}^{-3}
\end{aligned}$$

and then verify whether any of them are ever < 0.

Using the AMS approximation, and $\log\left(1+\frac{s}{b}\right) \sim \frac{s}{b} - \frac{1}{2}\left(\frac{s}{b}\right)^2$ we get the following

$$\begin{aligned}
\partial_s^2 \text{AMS} &\sim \frac{1}{b+s}\frac{\sqrt{b}}{s} - \frac{b^{3/2}}{s^3}\frac{s^2}{b^2} \\
&\sim \frac{1}{sb^{3/2}} - \frac{1}{sb^{\frac{1}{2}}} \\
&< 0
\end{aligned}$$

Since in this regime of large b this term is negative, the Hessian cannot be positive definite and hence the metric is non-convex. If the Hessian is *concave* though, we can simply optimize $-$AMS, so we need to check the other terms.

$$\begin{aligned}\partial_b^2 \text{AMS} &\sim \frac{sb}{s+b}\frac{\sqrt{b}}{s} + \sqrt{b} - \frac{b^{3/2}}{s^3}\left(\frac{s}{b} - \frac{1}{2}\left(\frac{s}{b}\right)^2 + sb\right)^2 \\ &= \frac{b^{3/2}}{s+b} + \sqrt{b} - \frac{1}{s\sqrt{b}} - \frac{b^{7/2}}{s} - 2\frac{b^{3/2}}{s} + \sqrt{b} < 0\end{aligned}$$

$$\begin{aligned}\partial_{bs}^2 \text{AMS} &\sim \frac{\sqrt{b}}{s(s+b)} + \frac{b^{3/2}}{s} - \frac{b^{3/2}}{s^3}\left(\frac{s}{b} - \frac{1}{2}\left(\frac{s}{b}\right)^2\right)\left(\frac{s}{b} - \frac{1}{2}\left(\frac{s}{b}\right)^2 + sb\right) \\ &\sim \frac{1}{s\sqrt{b}}\left(1 - \frac{s}{b}\right) + \frac{b^{3/2}}{s} - \frac{1}{s\sqrt{b}} + \frac{1}{2}\sqrt{b} - \frac{b^{3/2}}{s} \\ &\sim -\frac{1}{b^{3/2}} + \frac{1}{2}\sqrt{b} > 0\end{aligned}$$

B Modelling

B.1 Newton's Method for Logistic Regression

From the conditional distribution under the logistic regression model for a single point $\mathbf{x} \in R^d$, $y \in \{+1, -1\}$ and coefficients $\boldsymbol{\beta} \in R^d$.

$$\begin{aligned}p(y \mid \mathbf{x}, \boldsymbol{\beta}) &= \sigma(f(\mathbf{x}; \boldsymbol{\beta}) \cdot y) \\ &= (1 + \exp(y\langle \mathbf{x}, \boldsymbol{\beta}\rangle))^{-1}\end{aligned}$$

We want to minimise the negative loglikelihood:

$$\begin{aligned}l(\mathbf{x}, y, \boldsymbol{\beta}) &= \log(1 + \mathrm{e}^{y\langle \mathbf{x}, \mathbf{x}\rangle}) \\ \partial_{\beta_a} l &= \frac{y x_a \mathrm{e}^{y\langle \mathbf{x}, \boldsymbol{\beta}\rangle}}{1 + \mathrm{e}^{y\langle \mathbf{x}, \boldsymbol{\beta}\rangle}} \\ \partial^2_{\beta_a \beta_b} l &= \frac{y^2 x_a x_b \mathrm{e}^{y\langle \mathbf{x}, \boldsymbol{\beta}\rangle}}{(1 + \mathrm{e}^{y\langle \mathbf{x}, \boldsymbol{\beta}\rangle})^2}\left(1 + \mathrm{e}^{y\langle \mathbf{x}, \boldsymbol{\beta}\rangle} - \mathrm{e}^{y\langle \mathbf{x}, \boldsymbol{\beta}\rangle}\right) \\ &= \frac{x_a x_b \mathrm{e}^{y\langle \mathbf{x}, \boldsymbol{\beta}\rangle}}{(1 + \mathrm{e}^{y\langle \mathbf{x}, \boldsymbol{\beta}\rangle})^2} \\ &= x_a x_b w(x)\end{aligned}$$

where in the last line we define a weight function $w(x) = \frac{\mathrm{e}^x}{(1+\mathrm{e}^x)^2}$.

To implement Newton's method we want to try and simplify the expressions. First define the logistic and logit functions as

$$\begin{aligned}\gamma(x) &= (1 + \mathrm{e}^{-x})^{-1} \\ \gamma^{-1}(p) &= \log\left(\frac{p}{1-p}\right)\end{aligned}$$

and then write out the derivatives in terms of the logistic function:

$$\begin{aligned}\partial_{\beta_a} l(y = +1) &= x_a \gamma(\langle \mathbf{x}, \boldsymbol{\beta}\rangle) \\ \partial_{\beta_a} l(y = -1) &= -x_a(1 - \gamma(\langle \mathbf{x}, \boldsymbol{\beta}\rangle)\end{aligned}$$

If we then map $y \in \{1, -1\}$ to $y' \in \{0, 1\}$ we can condense these into a single expression

$$\partial_{\beta_a} l(y) = -x_a(y - \gamma(\langle \mathbf{x}, \boldsymbol{\beta} \rangle))$$

and rewrite the loss as $l(y') = \log(1 + \mathrm{e}^{(-1)^{y'} \langle \mathbf{x}, \boldsymbol{\beta} \rangle})$.

Now for our entire dataset D, assumed to be IID, we have a total loss, gradient and Hessian

$$\begin{aligned} L &= \sum_{i \in D} l_i = \sum_{i \in D} \log(1 + \mathrm{e}^{y_i \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}) \\ \partial_{\boldsymbol{\beta}} L &= -\sum_{i \in D} \mathbf{x}_i (y_i - \gamma(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)) \\ &= -\mathbf{X}^T (\mathbf{y} - \gamma(\mathbf{X}\boldsymbol{\beta})) \\ \boldsymbol{H} &= \sum_{i \in D} \mathbf{x}_i \mathbf{x}_i^T w(\mathbf{x}_i) \\ &= \mathbf{X}^T \boldsymbol{W} \mathbf{X} \end{aligned}$$

with $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \in R^{n \times d}$, $\boldsymbol{W} = \mathrm{diag}(w(\mathbf{x})) \in R^{n \times n}$.

B.1.1 Regularisation

We can add constraints to our minimisation problem, to help constrain the coefficients and effectively implement L_2 regularisation by formulating the following convex optimisation problem.

$$\begin{aligned} \min . \quad & L \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|^2 \leq c \end{aligned}$$

From this we can construct the Lagrangian $\mathcal{L} = L + \lambda(\|\boldsymbol{\beta}\|^2 - c)$ and hence augment the gradient and Hessian we calculated above as

$$\begin{aligned} \partial_{\boldsymbol{\beta}} \mathcal{L} &= \partial_{\boldsymbol{\beta}} L + 2\lambda\boldsymbol{\beta} \\ \tilde{\boldsymbol{H}} &= \boldsymbol{H} + 2\lambda \mathbf{I} \end{aligned}$$

Since we have a free parameter c that determines the optimal value $\lambda^\star(c)$ of our dual variable, we can equivalently ignore c and treat λ as a parameter to be tuned.

B.1.2 Newton's method

Algorithm 1: Newton's method

1. Initialise parameters; $\boldsymbol{\beta}_0$, λ_0, ϵ, $L_0 = L(\boldsymbol{\beta}_0)$, m
2. **for** *i=1:m* **do**
   - Compute $\partial_{\boldsymbol{\beta}} L$ and $\boldsymbol{H}$
   - Calculate Newton step $\Delta\boldsymbol{\beta} = -\boldsymbol{H}^{-1} \partial_{\boldsymbol{\beta}} L$
   - Compute new loss L
   - Test stopping criterion: $|L - L_0| < \epsilon$
3. Return $\hat{\boldsymbol{\beta}}$.