Abstract

The binomial deviance and the SVM hinge loss functions are two of the most widely used loss functions in machine learning. While there are many similarities between them, they also have their own strengths when dealing with different types of data. In this work, we introduce a new exponential family based on a convex relaxation of the hinge loss function using softness and class-separation parameters. This new family, denoted Soft-SVM, allows us to prescribe a generalized linear model that effectively bridges between logistic regression and SVM classification. This new model is interpretable and avoids data separability issues, attaining good fitting and predictive performance by automatically adjusting for data label separability via the softness parameter. These results are confirmed empirically through simulations and case studies as we compare regularized logistic, SVM, and Soft-SVM regressions and conclude that the proposed model performs well in terms of both classification and prediction errors.

1 Introduction

Binary classification has a long history in supervised learning, arising in multiple applications to practically all domains of science and motivating the development of many methods, including logistic regression, $k$-nearest neighbors, decision trees, and support vector machines [5]. Logistic regression, in particular, provides a useful statistical framework for this class of problems, prescribing a parametric formulation in terms of features and yielding interpretable results [7]. It suffers, however, from “complete separation” data issues, that is, when the classes are separated by a hyperplane in feature space. Support vector machines (SVMs), in contrast, aim at finding such a separating hyperplane that maximizes its margin to a subset of observations, the support vectors [9, 8]. While SVMs are robust to complete separation issues, they are not directly interpretable [6]; for instance, logistic regression provides class probabilities, but SVMs rely on post-processing to compute them, often using Platt scaling, that is, adopting a logistic transformation.

Both SVM and logistic regression, as usual in statistical learning methods, can be expressed as optimization problems with a data fitting loss function and a model complexity penalty loss. In Section 2, we adopt this formulation to establish a new loss function, controlled by a convex relaxation “softness” parameter, that comprehends both methods. Adopting then a generalized linear model formulation, we further explore this new loss to propose a new exponential family, Soft-SVM. Given that this new regularized regression setup features two extra parameters, we expect it to be more flexible and robust, in effect bridging the performance from SVM and logistic regressions and addressing their drawbacks. We empirically validate this assessment in Section 3 with a simulation study and case studies. We show, in particular, that the new method performs well, comparably to the best of SVM and logistic regressions and with a similar smaller computational cost of logistic regression.
where $\delta$ works together with the softness parameter $\kappa$, Note that $p_\kappa(x) \rightarrow x_+$ as $\kappa \rightarrow \infty$. Motivated by the property of both logistic and SVM cumulants that $b(\theta) = \theta + b(-\theta)$, we define the Soft-SVM canonical parameters and cumulant as

$$\theta_i = p_\kappa(\eta_i + \delta) - p_\kappa(\eta_i - \delta) \equiv f_{\kappa,\delta}(\eta_i) \quad \text{and} \quad b_{\kappa,\delta}(\theta) = \frac{1}{2} [p_\kappa(\theta + 2\delta) + p_\kappa(\theta - 2\delta)],$$

where $\delta$ is a separation parameter in the cumulant. Parameter $\delta$ works together with the softness parameter $\kappa$ to effective bridge between logistic and SVM regressions: when $\kappa = 1$ and $\delta = 0$, we have logistic regression, while as $\kappa \rightarrow \infty$ and $\delta \rightarrow 1$ we obtain SVM regression. This way, $b_{1,0} \equiv b$ and $b_{\kappa,\delta} \rightarrow s$ pointwise as $\kappa \rightarrow \infty$ and $\delta \rightarrow 1$. The Soft-SVM penalized log-likelihood is then

$$\ell_{\kappa,\delta}(\theta; y) = \sum_{i=1}^n y_i \theta_i - b_{\kappa,\delta}(\theta_i) - \frac{\lambda}{2} \| \beta \|_2^2 = \sum_{i=1}^n y_i \theta_i - b_{\kappa,\delta}(\theta_i) - \frac{\lambda}{2} \left[ \begin{array} {c} [\beta_0] \\ \beta \end{array} \right]^\top \left[ \begin{array} {cc} 0 & I_f \\ \beta & 0 \end{array} \right] \left[ \begin{array} {c} [\beta_0] \\ \beta \end{array} \right],$$

Figure 1: Cumulant (left) and mean (right) functions for logistic, SVM, and Soft-SVM(5, 0.8).
Figure 2: Plot of variance functions when $\kappa \in \{1, 1.5, 2, 5\}$ and $\delta = 1 - 1/\kappa = \{0, 1/3, 1/2, 4/5\}$ respectively.

where we make clear that the bias $\beta_0$ is not penalized when using the precision matrix $0 \oplus I_f$.

While the separation parameter $\delta$ has a clear interpretation from the cumulant function, we can also adopt an equivalent alternative parameterization in terms of a scaled separation parameter $\alpha = \kappa \delta$ since the shifted soft plus can be written as

$$p_\kappa(x + z\delta) = \frac{1}{\kappa} \log \left( 1 + e^{\kappa(x + z\delta)} \right) = \frac{1}{\kappa} \log \left( 1 + e^{\kappa x + z\alpha} \right).$$

This alternative parameterization is more useful when we discuss the variance function in Section 2.2.

The mean function corresponding to $b_{\kappa,\delta}$ is its first derivative,

$$\mu(\theta) = b'_{\kappa,\delta}(\theta) = \frac{1}{2} \left[ \expit\{\kappa(\theta + 2\delta)\} + \expit\{\kappa(\theta - 2\delta)\} \right].$$

(4)

With $v(x) = \expit(x)(1 - \expit(x)) = e^x/(1 + e^x)^2$, the variance function is

$$V_{\kappa,\delta}(\mu) = b''_{\kappa,\delta}(b'_{\kappa,\delta}(\mu))^{-1} = \frac{\kappa}{2} \left[ v(\kappa(b'_{\kappa,\delta}(\mu))^{-1} + 2\delta)) + v(\kappa(b'_{\kappa,\delta}(\mu))^{-1} - 2\delta)) \right].$$

(5)

where the inverse mean function $[b'_{\kappa,\delta}(\mu)]^{-1}$ is given in the next section. Figure 2 shows the plot of variance function when $\kappa$ equals to different values and assumes the corresponding value of $\delta$ is a function of $\kappa$: $\delta = 1 - 1/\kappa$.

Finally, to relate the canonical parameters $\theta_i$ to the features, our link function $g$ can be found by solving the composite function $f_{\kappa,\delta}^{-1} \circ [b'_{\kappa,\delta}(\mu)]^{-1}$ (see Section 2.1 for details). Figure 1 compares the cumulant and mean functions for Bernoulli/logistic regression, SVM, and Soft-SVM with $\kappa = 5$ and $\delta = 0.8$. Note how the Soft-SVM functions match the SVM behavior for more extreme values of the canonical parameter $\theta$ and the logistic regression behavior for moderate values of $\theta$, achieving a good compromise as a function of the softness $\kappa$.

2.1 Soft-SVM Regression

Given our characterization of the Soft-SVM exponential family distribution, performing Soft-SVM regression is equivalent to fitting a regularized generalized linear model. To this end, we follow a cyclic gradient ascent procedure, alternating between an iteratively re-weighted least squares (IRLS) step for Fisher scoring [7] to update the coefficients $\beta$ given the softness $\kappa$, and then updating $\kappa$ and $\alpha$ given the current values of $\beta$ via a Newton-Raphson scoring step.
Algorithm 1: Soft-SVM Regression

\[ [b_{κ,δ}(μ)]^{-1} = \frac{1}{κ} \left\{ \frac{1}{2} \log \left( \frac{μ}{1-μ} \right) + \text{asinh}(e^{h_{κ,δ}(μ)}) \right\}, \]

with

\[ h_{κ,δ}(μ) = \log \cosh(2κδ) + \log \frac{|μ - 1/2|}{\sqrt{μ(1-μ)},} \]

and

\[ f_{κ,δ}(θ) = \frac{θ}{2} + \frac{1}{κ} \text{asinh}(e^{H_{κ,δ}(θ)}) \quad \text{with} \quad H_{κ,δ}(θ) = -κδ + \log \frac{|κ|}{2}. \]

Note that in all these expressions δ is always expressed in the product κδ, so it is immediate to use the alternative parameter α instead of δ.

Our algorithm is summarized in Algorithm 1. For notation simplicity, we represent both bias and coefficients in a single vector β and store all features in a design matrix X, including a column of ones for the bias (intercept).

**Algorithm 1: Soft-SVM Regression**

- **Input**: Parameters
- **Parameters**: Penalty λ, convergence tolerance ε, and small perturbation 0 < ν < 1.
- **Output**: Estimated regression coefficients $\hat{β}$ (including bias $\hat{β}_0$), shape parameter $\hat{κ}$ and softness parameter $\hat{κ}$.
- **Initialization**: Set $α^{(0)} = 1, κ^{(0)} = 1$ and, for $i = 1, \ldots, n, μ_i^{(0)} = (y_i + ν)/(1 + 2ν)$ as a ν-perturbed version of y and $η_i^{(0)} = g_{κ,(θ)}(μ_i^{(0)})$.
- **Cyclic scoring iteration**: for $t = 0, 1, \ldots$ (until convergence) do
  - **κ-step**: Update $κ^{(t+1)} = κ^{(t)} - [∂^2ℓ_{κ,α}/∂κ^2]^{-1}(∂ℓ_{κ,α}/∂κ)$ via Newton’s method;
  - **α-step**: Update $α^{(t+1)} = α^{(t)} - \sum_{i=1}^{N} [y_i f'_{κ,α}(η_i) - b'_{κ,α}(θ_i)]/\sum_{i=1}^{N} [y_i f''_{κ,α} - b''_{κ,α}]$ via Newton’s method;
  - **β-step**: Compute weights $w_i = y_i f''_{κ,α}(η_i) - f''_{κ,α}(η_i)b'_{κ,α}(θ_i) - (f'_{κ,α}(η_i))2b''_{κ,α}(θ_i)$ and set $W^{(t)} = \text{Diag}_{i=1,\ldots,n} \{ w_i^{(t)} \}$; Update $β^{(t+1)}$ by solving $\left(X^TW^{(t)}X + λ(0\oplus I_f)\right)β^{(t+1)} = X^TW^{(t)}η^{(t)} - X^T\text{Diag}_{i=1,\ldots,n} \{ f'_{κ,α}(η_i) \}(y - μ(β^{(t)}))$;
  - Set $η^{(t+1)} = Xβ^{(t+1)}$, $μ(β^{(t)}) = b'_{κ}(θ)$, and $θ = [θ_1, \ldots, θ_n]^T$;
  - if $|ℓ_{κ,α}^{(t+1)}(η^{(t+1)}; y) - ℓ_{κ,α}^{(t)}(η^{(t)}; y)|/|ℓ_{κ,α}^{(t)}(η^{(t)}; y)| < ε$ then break
- return $β^{(t)}$, $α^{(t)}$ and $κ^{(t)}$.

### 2.2 Soft-SVM Variance Function

Since $h_{κ,δ}(μ) = u_{κ}(μ) = \log \cosh(2κ) + \log |μ - 1/2|/\sqrt{μ(1-μ)}$, we have that

\[ κ([b_{κ,δ}(μ)]^{-1} ± 2δ) = \frac{1}{2} \log \left( \frac{μ}{1-μ} \right) + \text{asinh}(e^{h_{κ,δ}(μ)}) ± 2α \]

and so the variance function in (5) can be written as $V_{κ,α}(μ) = κ \cdot r_α(μ)$ where, with $u_α(μ) = \logit(μ)/2 + \text{asinh}(\exp(h_α(μ)))$,

\[ r_α(μ) = \frac{1}{2} \left\{ v(u_α(μ) + 2α) + v(u_α(μ) - 2α) \right\}, \]
Figure 3: Top row has variance-mean plots when scale parameter $\kappa$ is fixed to 1 and $\alpha \in \{0, \alpha^*, 1, 2, 5\}$, where $\alpha^* \approx 0.66$ is the critical value when the peaks start separating. Bottom row has histograms for the corresponding distribution of mean values.

This way, parameters $\kappa$ and $\alpha$ have specific effects on the variance function: the softness parameter $\kappa$ acts as a dispersion, capturing the scale of $V(\mu)$, while the scaled separation $\alpha$ controls the shape of $V(\mu)$.

Figure 3 shows variance function plots for the ESL mixture dataset discussed in Section 2.4 when penalty parameter $\lambda$ and scale-parameter $\kappa$ are both fixed at 1, and only allow the shape parameter $\alpha$ to increase from 0 to 5. We can see that as $\alpha$ increases, the overall variance decreases and the plot gradually splits from a unimodal shaped curve into a bimodal shaped curve. In practice, we observe that as $\alpha$ increases to achieve higher mode separation, most observations are pushed to have fitted $\mu$ values around 0.5 and variance weights close to zero, caught in-between variance peaks.

We can then classify observations with respect to their mean and variance values: points with high variance weights have a large influence on the fit and thus on the decision boundary, behaving like “soft” support vectors; points with low variance weights and mean close to 0.5 are close to the boundary but do not influence it, belonging thus to a “dead zone” for the fit; finally, points with low variance weights and large max\{$\mu, 1-\mu$} are inliers and have also low influence in the fit. This classification is illustrated in Figure 4.

2.3 Handling numerical issues

As $\kappa$ increases we should anticipate numerical issues as $b_\kappa$ and $f_\kappa$ becomes non-differentiable up to machine precision. For this reason, we need to adopt numerically stable versions for asinh($e^x$), log cosh($x$) and log sinh($x$), needed by $g_\kappa$ and thus also by $V_\kappa$ above.

The first step is a numerically stable version for the Bernoulli cumulant; with $\epsilon$ the machine precision, we adopt

$$
\log 1pe(x) = \log(1 + e^x) = \begin{cases} 
  x, & \text{if } x > -\log \epsilon \\
  x + \log(1 + e^{-x}), & \text{if } 0 < x \leq -\log \epsilon \\
  \log(1 + e^x), & \text{if } \log \epsilon \leq x \leq 0 \\
  0, & \text{if } x < \log \epsilon.
\end{cases}
$$

This way, the soft plus can be reliably computed as $p_\kappa(x) = \log 1pe(\kappa x)/\kappa$. Again exploiting domain partitions and the parity of cosh, we can define

$$
\log \cosh(x) = \begin{cases} 
  |x| - \log 2, & \text{if } |x| > -\log \sqrt{\epsilon} \\
  |x| - \log 2 + \log 1pe(-2x), & \text{if } |x| \leq -\log \sqrt{\epsilon}.
\end{cases}
$$
Figure 4: Plots of classification results by using soft-SVM when scale parameter $\kappa$ is fixed to 1 and $\alpha$ equals to 0, 1, 1.2, 1.5, 2, and 5. The colors represent a gradient from blue ($\hat{\mu}_i = 0$) to orange ($\hat{\mu}_i = 1$). Filled dots have $V(\hat{\mu}_i) \geq 1$ while hollow dots have $V(\hat{\mu}_i) < 1$.

We branch twice for $\text{asinh}(e^x)$: we first define

$$\gamma(x) = \begin{cases} 1, & \text{if } |x| > -\log \sqrt{\epsilon} \\ \frac{1}{\sqrt{1 + e^{-2|x|}}}, & \text{if } |x| \leq -\log \sqrt{\epsilon} \end{cases}$$

and then

$$\text{asinh}(e^x) = \begin{cases} x + \log(1 + \gamma(x)), & \text{if } x > 0 \\ \log(e^x + \gamma(x)), & \text{if } x \leq 0 \end{cases}$$

Finally, we write

$$\log \sinh(x) = \begin{cases} \log \frac{1}{2} + x + \log(1 - e^{-2x}), & \text{if } x > 0 \\ \log \frac{1}{2} - x + \log(1 + e^{-2x}), & \text{if } x \leq 0 \end{cases}$$

to make it stable whenever $x$ is greater or smaller than 0.

2.4 Practical example: ESL mixture

To examine how the Soft-SVM regression works for classification in comparison with SVM and logistic regression, we carry out a detailed example using the simulated mixture data in [5], consisting of two features, $X_1$ and $X_2$. To estimate the complexity penalty $\lambda$ we use ten-fold cross-validation with 20 replications for all methods. The results are summarized in Figure 5. Because $\hat{\kappa}$ is not much larger than unit, the fitted curve resembles a logistic curve; however, the softness effect can be seen in the right panel as points with the soft margin have estimated probabilities close to 0.5, with fitted values quickly increasing as points are farther away from the decision boundary. To classify, we simply take a consensus rule, $\hat{y}_i = I(\hat{\mu}_i > 0.5)$, with $I$ being the indicator function.

Figure 6 shows the effect of regularization in the cross-validation performance of Soft-SVM regression, as measured by Matthews correlation coefficient (MCC). We adopt MCC as a metric because it achieves a more balanced evaluation of all possible outcomes [3]. As can be seen in the right panel, comparing MCC with SVM and logistic regressions, the performances are quite comparable; Soft-SVM regression seems to have a slight advantage on average, however, and logistic regression has a more variable performance.
3 Experiments

3.1 Simulation study

Prior works have shown that the performance of SVM can be severely affected when applied to unbalanced datasets since the margin obtained is biased towards the minority class \([1, 2]\). In this section, we conduct a simulation study to investigate these effects on Soft-SVM regression.

We consider a simple data simulation setup with two features. There are \(n\) observations, partitioned into two sets with sizes \(n_1 = \lfloor \rho n \rfloor\) and \(n_2 = n - n_1\), for an imbalance parameter \(0 < \rho \leq 0.5\). We expect to see SVM performing worse as \(\rho\) gets smaller and the sets become more unbalanced.

Observations in the first set have labels \(y_i = 0\) and \(x_i \sim N(\mu_1, \sigma^2 I_2)\), while the \(n_2\) observations in the second set have \(y_i = 1\) and \(x_i \sim N(\mu_2, \sigma^2 I_2)\). We set \(\mu_1 = (\sqrt{2}, 1)\) and \(\mu_2 = (0, 1 + \sqrt{2})\), so that the decision boundary is \(X_2 = X_1 + 1\) and the distance from either \(\mu_1\) or \(\mu_2\) to the boundary is 1. Parameter \(\sigma\) acts as a measure of separability: for \(\sigma < 1\) we should have a stronger performance.

Figure 5: Soft-SVM regression on ESL mixture data: fitted values \(\hat{y}\) (left) and decision boundary in feature space (right). Points are shaped by predicted classification, colored by observed labels, and have transparency proportional to \(|\hat{y}_i - 0.5|\). Hashed lines mark \(\delta\) in the fitted value plot and the soft margin \(M = \delta/\|\beta\|_2\) in the feature space plot. \(\delta = \alpha/\kappa\).

Figure 6: Cross-validation results (replication averages) for Soft-SVM using MCC (left) and performance comparisons to SVM and logistic regression (right).
from the classification methods, but as $\sigma$ becomes too small we expect to see complete separability challenging logistic regression.

3.2 Case studies

To evaluate the empirical performance of Soft-SVM regression and classification, we select nine well-known datasets from the UCI machine learning repository with a varied number of observations and features. Table lists the datasets and their attributes. Not all datasets are originally meant for binary classification; dataset Abalone, for instance, has three classes—male, female and infant—so we only take male and female observations and exclude infants. The red/white wine quality datasets originally have ten quality levels, labeled from low to high. We dichotomize these levels into a low quality class (quality levels between 1 and 5, inclusive) and a high quality class (quality levels 6 to 10). All remaining datasets have two classes.

As in the simulation study, we assess performance using MCC from 50 ten-fold cross-validation replications. For each replication, we estimate penalties $\lambda$ for all classification methods again using ten-fold cross-validations. Figure shows the results.
Table 1: Selected case study datasets from UCI repository.

| UCI dataset                              | # of observations \( n \) | # of features \( f \) |
|------------------------------------------|--------------------------|-----------------------|
| Abalone (only F & M classes)            | 4177                     | 8                     |
| Australian credit                       | 690                      | 14                    |
| Breast cancer                           | 569                      | 30                    |
| Haberman                                 | 306                      | 3                     |
| Heart disease                            | 303                      | 13                    |
| Liver disorder                           | 345                      | 6                     |
| Pima Indians diabetes                    | 768                      | 8                     |
| Red wine quality (amended)               | 1599                     | 11                    |
| White wine quality (amended)             | 4898                     | 11                    |

Figure 8: Performance comparison of MCC for Soft-SVM, SVM, and logistic regressions for nine UCI datasets over 50 ten-fold cross validation studies.

In general, it can be seen that SVM classification performs worse on “tough” datasets with low MCC performance, such as Abalone and Haberman, probably due to poor selection of penalty parameters. On the other hand, SVM performs favorably when compared to Soft-SVM and logistic regressions in the Heart dataset, being more robust to cases where many features seem to be non-informative of classification labels. Overall, Soft-SVM regression performs comparably to the best of SVM and logistic regression even in hard datasets, as it was observed in the simulation study. Interestingly, it has a clear advantage over both SVM and logistic regressions in cases where a few features are very informative, such as in the Abalone and Australian credit datasets. It often has an edge over logistic regression in over-dispersed datasets such as Breast cancer and White wine datasets. It seems that Soft-SVM regression can better accommodate these cases via higher dispersion on moderate values of softness, i.e. \( \kappa < 2 \), or via better use of regularization for increased robustness and generalizability.
4 Conclusion

We have presented a new binary classification and regression method based on a new exponential family distribution, Soft-SVM. Soft-SVM arises from extending a convex relaxation of the SVM hinge loss with a softness parameter, in effect achieving a smooth transformation between the binomial deviance and SVM hinge losses. Soft-SVM regression then entails fitting a generalized linear model, and so it is comparable to logistic regression in terms of computational complexity, having just a small overhead to fit the softness parameter. Moreover, since many points might have low variance weights by not being “soft” support vectors—either being dead-zoners or inliers, by our classification in Section 2.2—they can be selected out of the fit to achieve significant computational savings.

More importantly, as we have shown in a simulation study and multiple case studies, Soft-SVM classification performs comparably to the best of SVM and logistic regressions, often over performing them, even if slightly, in feature-rich datasets. This advantage in performance can be attributed to higher robustness from the softness and separation parameters, enabling the model to accommodate well over-dispersed datasets and adequately address separability issues via regularization. By bridging the SVM hinge and binomial deviance loss functions, it seems that Soft-SVM regression is able to combine robustness and interpretability, the advantages of SVM and logistic regression, respectively, freeing the practitioner from the burden of selecting one or the other.

For future work, we intend to develop data sampling procedures and adopt a Bayesian formulation to then sample posterior estimates of coefficients, softness and separation parameters. Another natural direction of work is to incorporate kernels, as it is common in SVM regression. Finally, we also expect to extend this formulation to multinomial classification [10], hoping again to bring some robustness from SVM while keeping interpretability.

References

[1] Christoph Bernau, Markus Riester, Anne-Laure Boulesteix, Giovanni Parmigiani, Curtis Hutenhower, Levi Waldron, and Lorenzo Trippa. Cross-study validation for the assessment of prediction algorithms. Bioinformatics, 30(12):i105–i112, 2014.

[2] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408:189–215, 2020.

[3] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics, 21(1):1–13, 2020.

[4] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, second edition, 2009.

[6] Yi Lin. Support vector machines and the bayes rule in classification. Data Mining and Knowledge Discovery, 6(3):259–275, 2002.

[7] P. McCullagh and J. A. Nelder. Generalized Linear Models. Chapman & Hall, second edition, 1989.

[8] Javier M Moguerza, Alberto Muñoz, et al. Support vector machines with applications. Statistical Science, 21(3):322–336, 2006.

[9] Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer, 2013.

[10] Ji Zhu and Trevor Hastie. Kernel logistic regression and the import vector machine. Journal of Computational and Graphical Statistics, 14(1):185–205, 2005.
A Appendix

A.1 $[b'_\kappa(\mu)]^{-1}$

\[
[b'_\kappa(\mu)]^{-1} = \frac{1}{\kappa} \log \left\{ \frac{-(2\mu - 1)(e^{-2\kappa\delta(\kappa)} + e^{2\kappa\delta(\kappa)})}{2(2\mu - 2)} + \sqrt{\left[ \frac{-(2\mu - 1)(e^{-2\kappa\delta(\kappa)} + e^{2\kappa\delta(\kappa)})}{2(2\mu - 2)} \right]^2 - \frac{\mu}{1 - \mu}} \right\}
\]

\[
= \frac{1}{\kappa} \log \left\{ \frac{\sqrt{\frac{\mu}{1 - \mu}} \left( \frac{\mu}{1 - \mu} \frac{1}{2} \left( e^{-2\kappa\delta(\kappa)} + e^{2\kappa\delta(\kappa)} \right) \right)}{\sqrt{\mu(1 - \mu)}} \right\}
\]

\[
= \left\{ \begin{array}{ll}
\frac{\theta}{\kappa} + \frac{1}{\kappa} \log(e^{H_\kappa(\theta)} + \sqrt{e^{2H_\kappa(\theta)} + 1}), & \text{if } \theta > 0 \\
\frac{\theta}{2} + \frac{1}{\kappa} \log(-e^{H_\kappa(\theta)} + \sqrt{e^{2H_\kappa(\theta)} + 1}), & \text{if } \theta \leq 0
\end{array} \right.
\]

\[
h_\kappa(\mu) = \log \left\{ \frac{|\mu - \frac{1}{2}|}{\sqrt{\mu(1 - \mu)}} \cdot \frac{1}{2} (e^{-2\kappa\delta(\kappa)} + e^{2\kappa\delta(\kappa)}) \right\} = \log \cosh(2\kappa\delta(\kappa)) + \log \frac{|\mu - \frac{1}{2}|}{\sqrt{\mu(1 - \mu)}}.
\]

A.2 $f^-1_\kappa(\theta)$

\[
f^-1_\kappa(\theta) = \frac{1}{\kappa} \log \left\{ \frac{e^{\kappa\theta}}{2e^{\kappa\delta(\kappa)}} + \sqrt{e^{\kappa\theta} + \left( \frac{e^{\kappa\theta} - 1}{2e^{\kappa\delta(\kappa)}} \right)^2} \right\}
\]

\[
= \frac{1}{\kappa} \log \left\{ \frac{e^{\kappa\theta}}{2} + \frac{e^{\kappa\theta} - 1}{2e^{\kappa\delta(\kappa)}} + \sqrt{1 + \left( \frac{e^{\kappa\theta} - 1}{2e^{\kappa\delta(\kappa)}} \right)^2} \right\}
\]

\[
= \frac{1}{\kappa} \log \left\{ \frac{e^{\kappa\theta}}{2} + \frac{1}{\kappa} \log \left\{ \frac{e^{\kappa\theta} - 1}{2e^{\kappa\delta(\kappa)}} + \sqrt{1 + \left( \frac{e^{\kappa\theta} - 1}{2e^{\kappa\delta(\kappa)}} \right)^2} \right\} \right\}
\]

\[
= \left\{ \begin{array}{ll}
\frac{\theta}{2} + \frac{1}{\kappa} \log(e^{H_\kappa(\theta)} + \sqrt{e^{2H_\kappa(\theta)} + 1}), & \text{if } \theta > 0 \\
\frac{\theta}{2} + \frac{1}{\kappa} \log(-e^{H_\kappa(\theta)} + \sqrt{e^{2H_\kappa(\theta)} + 1}), & \text{if } \theta \leq 0
\end{array} \right.
\]

\[
H_\kappa(\theta) = \log \left\{ e^{-\kappa\delta(\kappa)} \cdot \frac{|e^{\kappa\theta} - 1|}{2(e^{\kappa\theta})^2} \right\}
\]

\[
= -\kappa\delta(\kappa) + \log \sinh(\frac{\kappa|\theta|}{2}),
\]

\[
\log \sinh(x) = \log \left( \frac{e^{2x} - 1}{2e^x} \right) = \left\{ \begin{array}{ll}
\log \frac{1}{2} + x + \log(1 - e^{-2x}), & \text{if } x > 0 \\
\log \frac{1}{2} - x + \log(1 - e^{2x}), & \text{if } x \leq 0
\end{array} \right.
\]

A.3 $V_\kappa(\mu)$

\[
V_\kappa(\mu) = b''_\kappa(\theta) = b''_\kappa(g_\kappa(\mu)),
\]
\[ b''_n(\theta) = \frac{\partial^2}{\partial^2 \theta} \left\{ 0.5 * \left[ \frac{\log(1+e^{(\theta-2\delta(n))})}{\kappa} + \log(1+e^{(\theta+2\delta(n))}) \right] \right\} \]

\[ = \frac{\partial}{\partial \theta} \left\{ 0.5 * \left[ \frac{e^{\kappa(\theta-2\delta(n))}}{1+e^{(\theta-2\delta(n))}} + \frac{e^{\kappa(\theta+2\delta(n))}}{1+e^{(\theta+2\delta(n))}} \right] \right\} \]

\[ = 0.5 * \left[ \frac{\kappa e^{\kappa(\theta-2\delta(n))}}{(1+e^{(\theta-2\delta(n))})^2} + \frac{\kappa e^{\kappa(\theta+2\delta(n))}}{(1+e^{(\theta+2\delta(n))})^2} \right] \]

\[ = \frac{\kappa}{2} \left\{ v[\kappa(\theta-2\delta(n))] + v[\kappa(\theta+2\delta(n))] \right\}, \]

\[ \delta(n) = 1 - \frac{1}{\kappa}, v(x) = \frac{e^x}{1+e^x}. \]

### A.4 \( \beta \)-step

\[ l = \sum_{i=1}^{n} y_i b_\theta(\theta_i), \theta_i = f(\eta_i), y_i = X_i^\top \beta = \beta_0 + x_i \beta_1 + \ldots + x_p \beta_p, \]

\[ U(\beta)_{j=0,1,...,p} = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial (y_i - b_\theta(\theta_i))}{\partial \theta_i} \cdot \beta_j \]

\[ U(\beta) = \frac{\partial l}{\partial \beta} = \begin{bmatrix} U(\beta_0) \\
U(\beta_1) \\
\vdots \\
U(\beta_p) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} f_n'(y_i) \cdot X_{i0} \cdot (y_i - b'_n(\theta_i)) \\
\sum_{i=1}^{n} f_n'(y_i) \cdot X_{i1} \cdot (y_i - b'_n(\theta_i)) \\
\vdots \\
\sum_{i=1}^{n} f_n'(y_i) \cdot X_{ip} \cdot (y_i - b'_n(\theta_i)) \end{bmatrix} = X^\top \text{Diag}_{i=1,...,n} \{f'_n(\eta_i)\} \begin{bmatrix} y_1 - b'_n(\theta_1) \\
y_2 - b'_n(\theta_2) \\
\vdots \\
y_n - b'_n(\theta_n) \end{bmatrix} \]

\[ = X^\top \text{Diag}_{i=1,...,n} \{f'_n(\eta_i)\}(y - \mu(\theta)), \text{where } \mu(\theta) = \begin{bmatrix} b'_n(\theta_1) \\
b'_n(\theta_2) \\
\vdots \\
b'_n(\theta_n) \end{bmatrix}. \]

\[ H(\beta) = \frac{\partial U(\beta)}{\partial \beta^\top} = \begin{bmatrix} \sum_{i=1}^{n} f_n'(y_i) \cdot X_{i0} \cdot (y_i - b'_n(\theta_i)) \\
\sum_{i=1}^{n} f_n'(y_i) \cdot X_{i1} \cdot (y_i - b'_n(\theta_i)) \\
\vdots \\
\sum_{i=1}^{n} f_n'(y_i) \cdot X_{ip} \cdot (y_i - b'_n(\theta_i)) \end{bmatrix} \]

For each row \( j, j = 0, 1, 2, ..., p: \)

\[ \frac{\partial}{\partial \beta} \sum_{i=1}^{n} f_n'(y_i) \cdot X_{ij} \cdot (y_i - b'_n(\theta_i)) = \frac{\partial}{\partial \beta^\top} \sum_{i=1}^{n} f_n'(y_i) X_{ij} y_i + \frac{\partial}{\partial \beta} \sum_{i=1}^{n} f_n'(y_i) X_{ij} b'_n(\theta_i) \]

\[ \begin{bmatrix} \sum_{i=1}^{n} X_{ij} f_n''(y_i) - f_n''(\eta_i) b'_n(\theta_i) - (f_n'(y_i))^2 b''_n(\theta_i) \end{bmatrix}^\top \]

\[ = \begin{bmatrix} \sum_{i=1}^{n} X_{ij} f_n''(y_i) - f_n''(\eta_i) b'_n(\theta_i) - (f_n'(y_i))^2 b''_n(\theta_i) \end{bmatrix}^\top X_{i0} \]

\[ \cdots \]

\[ \sum_{i=1}^{n} X_{ij} f_n''(y_i) - f_n''(\eta_i) b'_n(\theta_i) - (f_n'(y_i))^2 b''_n(\theta_i) \]
Therefore,

\[
H(\beta) = \begin{bmatrix}
\sum_{i=1}^{n} X_i \beta_i & \cdots & \sum_{i=1}^{n} X_i \beta_i \\
\sum_{i=1}^{n} X_i \beta_i & \cdots & \sum_{i=1}^{n} X_i \beta_i \\
\vdots & \ddots & \vdots \\
\sum_{i=1}^{n} X_i \beta_i & \cdots & \sum_{i=1}^{n} X_i \beta_i
\end{bmatrix}^T
\]

\[
= X^T \text{Diag}_{i=1,...,n} \{ y_i f^{\prime\prime}_{\kappa}(\eta_i) - f^{\prime\prime}_{\kappa}(\eta_i) b_{\kappa}^\prime(\theta_i) - (f^{\prime\prime}_{\kappa}(\eta_i))^2 b_{\kappa}^\prime(\theta_i) \} \approx w_i \} X
\]

\[
= X^T (\text{Diag}_{i=1,...,n} \{ \omega_i \}) W(\beta) X
\]

\[
= X^T W(\beta^t) X.
\]

Since by Newton’s method we have:

\[
\beta^{(t+1)} \approx \beta^{(t)} - [H^{(t)}]^{-1} U(\beta^{(t)}) \Rightarrow H^{(t)} \beta^{(t+1)} \approx H^{(t)} \beta^{(t)} - U(\beta^{(t)}),
\]

and by adding the penalty term and plugging in the value of \( H \) and \( U \), we can update \( \beta^{(t+1)} \) by solving the following matrix equation:

\[
(X^T W^{(t)} X + \lambda(0 \otimes I_f)) \beta^{(t+1)} = X^T W^{(t)} \eta^{(t)} - X^T \text{Diag}_{i=1,...,n} \{ f^{\prime\prime}_{\kappa}(\eta_i) \}(y - \mu(\beta^{(t)})), \quad \eta^{(t)} = X \beta^{(t)}.
\]