Optimizing Binary Regression: A Comparative Analysis of Gradient Descent vs. Newton-Raphson and IRLS Across Link Functions,

Achraf Cherkaoui[1], Evrim Oral[1*]

[1]Louisiana State University Health Sciences Center , School of Public Health, Biostatistics and Data Science Program

LSU Health NEW ORLEANS School of Public Health

Introduction

While the logit link is the de facto standard for binary outcomes in machine learning (ML), the influence of alternative link functions on the optimization landscape remains under-documented. Modern ML predominantly relies on first-order Gradient Descent (GD) for scalability; however, the efficiency of these solvers can vary significantly depending on the choice of link. This study systematically compares first-order (GD) and second-order optimization methods—specifically Newton-Raphson and Iteratively Reweighted Least Squares (IRLS)—across four link functions: **Logit, Probit, Complementary Log-Log (Clog-log), and Cauchit.** We evaluate these based on three critical metrics: convergence rates, numerical stability (particularly near the boundaries of the probability space), and computational cost per iteration.

Methods

- **Data-Generating Process:**

Data was simulated to evaluate binary regression models under well-specified conditions:

- $P(Y_i|x_i) = p_i$ where $F^{-1}(p_i) = \sum_j \beta_{ij} x_{ij}$

- **Link Functions Evaluated (*F*):**

Logit (canonical), Probit, Complementary Log-Log (Cloglog), and Cauchit.

- **Optimizers Implemented in R:**

Gradient Descent (GD): First-order method with backtracking line search. Newton-Raphson (NR): Second-order method using the observed Hessian. Iteratively Reweighted Least Squares (IRLS): Second-order method using the expected Fisher information.

- **Convergence Criterion:** Optimization was halted when $\|\nabla \ell(\beta)\|_2 < 10^{-6}$ or at max iterations.
- **Number of covariates: 2**

Results

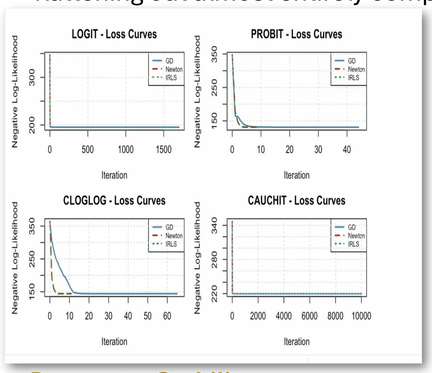
-The Computational Bottleneck:

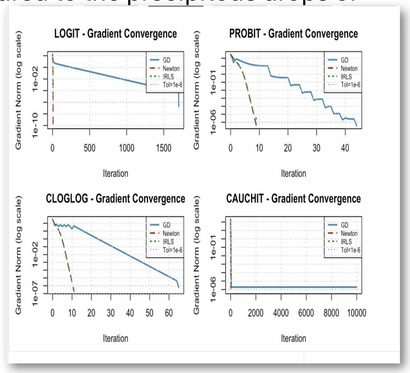
- **Second-Order Methods Yield Massive Speedups for Heavy-Tailed Links:** The computational cost of GD increases exponentially when moving away from the canonical logit link, whereas Newton-Raphson and IRLS remain highly efficient.
- **The Cauchit Anomaly:** GD required **10,000 iterations** (hitting the maximum threshold) to converge under the Cauchit link, compared to just 8 iterations for both Newton and IRLS.
- **Relative Speedup:** Newton-Raphson was roughly **20x faster** than GD for Logit, but a staggering **15055x faster** for Cauchit.

Results

-Convergence Trajectories:

- **Gradient Norm and Loss Curves Expose GD's Limitations:** Tracking the negative log-likelihood and gradient norm per iteration reveals why GD fails on non-canonical links.
- **Immediate vs. Asymptotic Convergence:** Both Newton and IRLS achieve near-instantaneous convergence (typically under 10 iterations) across all links. GD exhibits a slow, linear asymptotic crawl.
- **Gradient Stagnation:** The log-scale gradient convergence paths show GD struggling to find the minimum for the Cauchit link, flattening out almost entirely compared to the precipitous drops of





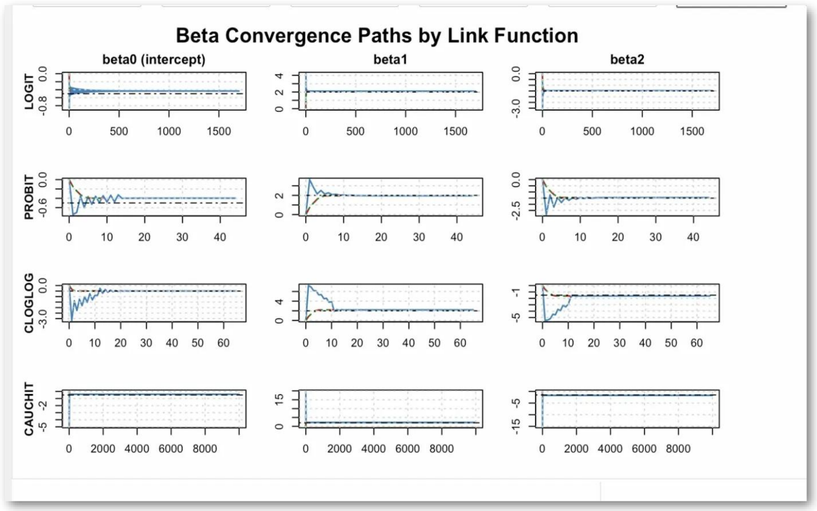
-Parameter Stability:

Estimating β:

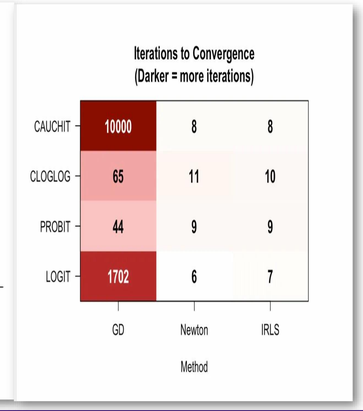
Monitoring the coefficient paths (β_0, β_1, β_2) across iterations highlights the precision of second-order scaling.

Newton and IRLS correct initial estimates rapidly, finding the true parameter values within a few iterations without significant oscillation.

GD exhibits significant drift and slow stabilization, particularly for the intercept (β_0) in the Cloglog and Probit models.




```
--- Total Time (milliseconds) ---

    Link GD_ms Newton_ms IRLS_ms        Fastest
   LOGIT  1342        67       87 Newton.elapsed
  PROBIT    58        65        6   IRLS.elapsed
 CLOGLOG    62         8       10 Newton.elapsed
 CAUCHIT 30110         2        3 Newton.elapsed

--- Time Per Iteration (milliseconds) ---

    Link GD_per_iter Newton_per_iter IRLS_per_iter
   LOGIT      0.7885         11.1667        12.4286
  PROBIT      1.3182          7.2222         0.6667
 CLOGLOG      0.9538          0.7273         1.0000
 CAUCHIT      3.0110          0.2500         0.3750

--- Speedup: How much faster is Newton/IRLS vs GD? ---

    Link Newton_speedup IRLS_speedup
   LOGIT           20.0         15.4
  PROBIT            0.9          9.7
 CLOGLOG            7.8          6.2
 CAUCHIT        15055.0      10036.7
```



Clinical Application Examples

Discrete-Time Survival & Infection Modeling (Cloglog Link): In clinical trials time-to-event outcomes is observed in discrete intervals between visits. First-order methods produce the same drifting β paths we observed, confirming that second-order optimization is required to prevent parameter oscillation when modeling rare clinical events.

Meta-Analyses & Extreme Clinical Heterogeneity (Cauchit Link): Aggregating data across multiple clinical trials can distort standard logistic models. The cauchit link provides heavy-tailed robustness against such extremes. Our finding that GD required 10,000 iterations (hitting the maximum) while Newton/IRLS converged in just 8 iterations serves as a direct cautionary tale against naive solvers in high-stakes, heavy-tailed medical models.

Method	Order	Medical Application	Algorithmic Behavior
GD	1st	Deep Learning & Large Bio-banks. Used when you have millions of rows (e.g., Genetic sequencing).	Scalable and simple. It only looks at the slope. It is robust but can be very slow to converge if the link function has a flat plateau
NR	2nd	Clinical Trials & Precision Medicine. Used when accuracy is more important than speed (e.g., small, high-stakes patient cohorts).	High precision. It looks at the slope and the curvature. It reaches the peak in very few steps but can crash (diverge) if the link function is too curvy (e.g. Cauchit)
IRLS	2nd	Epidemiological Risk Modeling. The standard for "Generalized Linear Models" (GLMs) in medical software like SAS or R.	A stabilized version of NR. It uses the Expected curvature (Fisher Information) rather than the observed curvature. This makes it much more stable than pure NR for messy medical data

Conclusion

Curvature is Critical: While GD is viable (though slow) for the canonical Logit link, it becomes practically unusable for heavy-tailed link functions like Cauchit due to flat regions in the log-likelihood surface.

Optimizer Selection: Second-order methods (NR and IRLS) are virtually immune to the scaling issues introduced by non-canonical links, converging reliably in under 12 iterations regardless of the link function.

Practical Recommendation: When fitting binary regression models with non-canonical links, second-order optimization (specifically IRLS/Fisher Scoring to guarantee positive semi-definiteness) is strictly required for timely and stable convergence.