# Beyond Maximum Likelihood: Variational Inequality Estimation for Generalized Linear Models

**Linglingzhi Zhu**
Georgia Tech
llzzhu@gatech.edu

**Jonghyeok Lee**
Georgia Tech
jlee4177@gatech.edu

**Yao Xie**
Georgia Tech
yao.xie@isye.gatech.edu

## Abstract

Generalized linear models (GLMs) are fundamental tools for statistical modeling, with maximum likelihood estimation (MLE) serving as the classical approach for parameter inference. While MLE performs well for canonical GLMs, it can become computationally challenging in more general settings with non-canonical, non-smooth, or saturating link functions, where the resulting optimization landscape may be ill-conditioned, non-convex, or non-differentiable. In this paper, we study an alternative estimation framework based on variational inequalities (VIs), which formulates GLM estimation through an operator-based equilibrium condition rather than likelihood minimization. We analyze the VI estimator from a statistical perspective and establish finite-sample error bounds and asymptotic normality, together with convergence guarantees for fixed-point and stochastic approximation algorithms. The framework accommodates a broad class of link functions, including non-canonical, non-smooth, and non-monotone cases satisfying a strong Minty-type condition, and extends naturally to generalized additive models (GAMs) via basis expansion. Numerical experiments demonstrate that under a fixed iteration budget, the VI approach achieves competitive estimation accuracy and faster, more stable convergence than MLE, particularly in GLMs and GAMs with non-canonical or non-smooth link functions.

## 1 Introduction

Generalized linear models (GLMs) are a cornerstone of statistics and machine learning, providing a unified framework that extends linear regression to accommodate diverse response types through the exponential-family formulation [20, 22]. By linking the conditional mean of the response to a linear predictor via a specified link function, GLMs combine interpretability and flexibility, encompassing classical models such as linear, logistic, and Poisson regression as special cases. Owing to this balance, GLMs have been widely used across scientific, engineering, and data-driven decision-making domains.

The central inferential task in GLMs is to estimate the regression parameter characterizing the relationship between covariates and responses. Maximum likelihood estimation (MLE) has long served as the standard approach, offering strong asymptotic guarantees such as consistency and efficiency under correct model specification. However, in modern modeling scenarios, MLE can face substantial computational challenges. When the link function departs from the canonical form, for example by incorporating saturating or non-smooth structures designed for robustness and flexibility, the resulting likelihood-based objective may become ill-conditioned, non-convex, or non-differentiable. These issues often lead to slow convergence, sensitivity to initialization, and numerical instability, particularly in high-dimensional settings. Such difficulties motivate alternative formulations that retain the interpretability of GLMs while improving computational behavior and theoretical robustness.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this paper, we study *variational inequality (VI)-based estimation* as an operator-based alternative to MLE for GLMs. The VI framework, with roots in optimization and operator theory [4] and close ties on the statistical and algorithmic sides to the generalized method of moments [7] and the perceptron [25], formulates estimation through an equilibrium condition rather than likelihood minimization, and has been applied to nonlinear least-squares and signal recovery problems [11]. This perspective naturally accommodates non-canonical and even non-monotone link functions, relaxing the smoothness and convexity requirements inherent in likelihood-based approaches. It also enables a unified treatment of statistical accuracy and algorithmic convergence within a single framework.

**Contributions.** Our analysis shows that the VI formulation recovers the MLE first-order optimality condition under canonical links and yields a novel moment-based estimator for non-canonical links. For the sample-average VI estimator, we establish finite-sample error bounds and asymptotic normality with a sandwich-type covariance structure under non-smooth link functions (Section 4), together with convergence guarantees for fixed-point and stochastic approximation algorithms (Section 5). By operating under a strong Minty condition rather than strong monotonicity or likelihood convexity, our framework accommodates non-canonical, saturating, and non-monotone links and extends naturally to semiparametric generalized additive models (GAMs, Section 6). Under correct specification, while converged MLE attains the Cramér–Rao lower bound when the global optimum is reachable, the VI estimator provides two key advantages in challenging non-canonical regimes: (i) convergence guarantees that MLE lacks when the likelihood is non-convex or has plateaus, and (ii) faster and more stable progress under a fixed first-order computational budget.

## 2 Background and Problem Setup

We briefly review the GLM framework and establish notation used throughout the paper. We then describe maximum likelihood estimation (MLE) and an alternative variational inequality formulation (VI), which forms the basis of our analysis. A summary of notation is provided in Appendix A.

Given a sample $(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in \mathbb{R}^d$ denotes the covariates and $y \in \mathbb{R}$ the response, a GLM specifies the conditional distribution of $y \mid \boldsymbol{x}$ using the exponential family. The conditional mean is linked to a linear predictor through a link function $g : \mathbb{R} \to \mathbb{R}$:

$$\mathbb{E}[y \mid \boldsymbol{x}] = g^{-1}\left(\beta_0 + \sum_{j=1}^{d} \beta_j x_j\right), \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d) \in \mathbb{R}^{d+1}$ is the coefficient vector including an intercept. Classical examples include logistic regression with a binomial distribution and logit link, and Poisson regression with a Poisson distribution and logarithmic link.

**Maximum likelihood estimation.** Given a conditional model with density (or mass) function $p(y \mid \boldsymbol{x}; \boldsymbol{\beta})$ and i.i.d. observations $\{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$ (where $N$ denotes the sample size), the MLE minimizes the empirical negative log-likelihood (NLL),

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} \mathcal{L}_N(\boldsymbol{\beta}) := \frac{1}{N} \sum_{i=1}^{N} \ell\left(g^{-1}\left(\beta_0 + \sum_{j=1}^{d} \beta_j x_j^i\right), y^i\right), \tag{MLE}$$

where $\ell(u, y)$ denotes the per-observation NLL expressed as a function of the mean $u$. Problem (MLE) can be solved using second-order methods such as Fisher scoring, IRLS, or quasi-Newton methods (e.g., L-BFGS) [6, 20], as well as first-order or stochastic methods in large-scale settings [17, 23].

When canonical link functions are used, (MLE) is convex and smooth, ensuring global optimality and efficient convergence. Some non-canonical links, such as the probit link, preserve this structure, but many links chosen in applications for physical or robustness reasons break it in one of three ways. First, saturating links create plateaus on which the NLL gradient vanishes. For example, photon-limited imaging constrains the Poisson intensity to a prescribed interval, which amounts to a clipped link [2]. Second, mixture and heavy-tailed links make the NLL non-convex, as in polychromatic X-ray tomography, where the measurements follow a mixture of exponential attenuation terms [19], and in robust binary regression with robit [18] or arctangent links. Third, links with linear rather than

exponential growth make the NLL ill-conditioned: the softplus link, used in Poisson models of neural spike trains [21], yields an NLL whose curvature decays for large predictors, and the identity link in point-process models yields gradients that are not Lipschitz near zero intensity [9]. Non-canonical links also appear in modern AI systems that fit GLMs on learned embeddings and model only the conditional mean, such as reward modeling from preference data, where the logistic link is not a necessary choice [26], and autoregressive sequence modeling [29].

**Variational inequality formulation.** We now formulate estimation as a variational inequality. Let $\mathbb{P}$ denote the joint distribution of $(\boldsymbol{x}, y)$ and let $V : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ denote the population operator

$$V(\boldsymbol{\beta}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[\left(g^{-1}\left(\beta_0 + \sum_{j=1}^{d}\beta_j x_j\right) - y\right)\tilde{\boldsymbol{x}}\right],$$

where $\tilde{\boldsymbol{x}} := [1; \boldsymbol{x}] \in \mathbb{R}^{d+1}$. We define the target parameter $\boldsymbol{\beta}^\star$ as the solution to the population variational inequality

$$\langle V(\boldsymbol{\beta}^\star),\, \boldsymbol{\beta} - \boldsymbol{\beta}^\star\rangle \geq 0, \qquad \forall\, \boldsymbol{\beta} \in \mathbb{R}^{d+1}. \tag{2}$$

If the univariate conditional mean relation (1) holds at $\boldsymbol{\beta}^\star$, the tower property of conditional expectation directly implies the $(d+1)$-dimensional orthogonality condition

$$V(\boldsymbol{\beta}^\star) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\big[\big(g^{-1}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}^\star) - y\big)\tilde{\boldsymbol{x}}\big] = \mathbb{E}_{\boldsymbol{x}}\big[\tilde{\boldsymbol{x}}\,\mathbb{E}\big[g^{-1}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}^\star) - y \mid \boldsymbol{x}\big]\big] = \mathbf{0}.$$

When a canonical link is used, (2) coincides with the population first-order optimality condition of the MLE. For non-canonical links, (2) defines a moment-based target parameter characterized solely by the conditional mean (1), removing the likelihood-specific derivative weight from the score equation. Given samples $\{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$, the empirical VI estimator $\hat{\boldsymbol{\beta}}_N$ is defined as

$$\hat{\boldsymbol{\beta}}_N \in \left\{\hat{\boldsymbol{\beta}} \in \mathbb{R}^{d+1} : \langle V_N(\hat{\boldsymbol{\beta}}),\, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle \geq 0,\ \forall\, \boldsymbol{\beta} \in \mathbb{R}^{d+1}\right\}, \tag{VI}$$

with $V_N(\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^N V_{(\boldsymbol{x}^i,y^i)}(\boldsymbol{\beta})$, where $V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}) := (g^{-1}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}) - y)\tilde{\boldsymbol{x}}$ denotes the per-sample vector field associated with a pair $(\boldsymbol{x}, y)$. Note that $V(\boldsymbol{\beta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}[V_{(\boldsymbol{x},y)}(\boldsymbol{\beta})]$.

**Relation to prior work.** VI-based estimation was introduced by Juditsky and Nemirovski [11], building on perceptron and isotonic methods such as Isotron and GLMtron [14, 15, 25], and was later extended to point processes, time series, and tomographic imaging [3, 12, 13, 19, 29]. Our work departs from this literature in both statistical scope and structural assumptions. First, whereas prior VI analyses focus primarily on algorithmic convergence, we study the estimator $\hat{\boldsymbol{\beta}}_N$ as a statistical object by establishing finite-sample bounds, asymptotic normality with a sandwich covariance under non-smooth links, and the explicit efficiency–stability trade-off relative to MLE. Second, while existing VI and matching-loss methods [11, 14, 15] rely on strong monotonicity, we operate under the strictly weaker *strong Minty condition*. Requiring only an overall growth trend toward the true parameter rather than pairwise monotonicity, this relaxation extends statistical consistency and linear fixed-point convergence to accommodate saturating plateaus (e.g., clipped links) and non-monotone links. Finally, while $\hat{\boldsymbol{\beta}}_N$ is a moment estimator [7], our finite-sample bounds and convergence guarantees do not follow from classical asymptotic generalized method of moments theory.

## 3 From MLE to VI

We consider the GLM defined in (1), where the response follows an exponential-family distribution with mean $g^{-1}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta})$. The choice of the loss $\ell$ and inverse link $g^{-1}$ significantly impacts the optimization landscape.

**Equivalence under canonical links.** When the model employs a canonical link function (e.g., Logit for Logistic, Identity for Normal), the gradient of the NLL coincides exactly with the VI operator: $\nabla\mathcal{L}_N(\boldsymbol{\beta}) = V_N(\boldsymbol{\beta})$. This implies that MLE and VI share identical first-order optimality conditions. Representative examples are summarized in Table 1. However, for non-canonical links, this equivalence breaks, and MLE often suffers from optimization challenges. Detailed derivations are provided in Appendix B.

Table 1: MLE and VI formulations under different links.

| Model | Loss $\ell(u, y)$ | Inverse Link $g^{-1}(z)$ | Equivalent? |
|---|---|---|---|
| Logistic | $-y \log u - (1-y)\log(1-u)$ | $e^z/(1+e^z)$ | ✓ |
| | | $\frac{1}{2} + \frac{1}{\pi}\arctan(z)$ | ✗ |
| Normal | $\frac{1}{2}(u-y)^2$ | $z$ | ✓ |
| | | $\max\{0, z\}$ | ✗ |
| Exponential | $\log u + y/u$ | $z^{-1}$ | ✓ |
| | | $e^{-z}$ | ✗ |
| Poisson | $-y \log u + u$ | $e^z$ | ✓ |
| | | $\log(1+e^z)$ | ✗ |
| | | $\max\{c, \min\{e^z, C\}\}$ | ✗ |

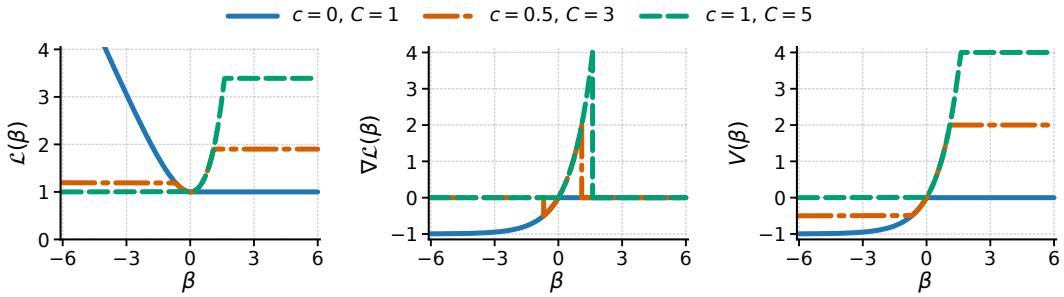



Figure 1: Clipped exponential link: MLE loss, its derivative, and the VI vector field.

**Flatness of MLE landscape.** It is well known that MLE optimization algorithms can converge slowly near the true parameter, especially in the non-canonical settings when the likelihood surface is nearly flat or ill-conditioned. Consider the softplus link $g^{-1}(z) = \log(1+e^z)$ in Poisson regression. In this case, the Hessian of the MLE objective decays rapidly in the right tail (as $z \to \infty$). In contrast, the Jacobian of the VI operator remains well-conditioned. This flatness in MLE leads to weaker local contraction and slower convergence compared to VI.

**Non-convexity and non-smoothness.** More severe issues arise with robust or truncated links. Consider the two-sided clipped exponential link

$$g^{-1}(z) = \max\{c, \min\{e^z, C\}\}, \qquad 0 < c < C,$$

and the Poisson negative log-likelihood

$$(\ell \circ g^{-1})(z) = -y \log(g^{-1}(z)) + g^{-1}(z).$$

This composition is piecewise smooth with flat regions for $z \le \log c$ and $z \ge \log C$, and is non-differentiable at the two kink points. A direct inspection of the one-sided derivatives shows that the monotonicity conditions required for convexity cannot hold simultaneously when $0 < c < C$. Hence, $\ell \circ g^{-1}$ is globally non-convex and non-smooth. For illustration, consider a single observation with $y = 1$ and predictor $z = \beta$, which gives

$$\mathcal{L}(\beta) = -\log g^{-1}(\beta) + g^{-1}(\beta).$$

As shown in the first two panels of Figure 1, clipping induces wide flat plateaus where gradient-based MLE updates can stall and fail to reach the global minimizer at $\beta = 0$. In contrast, the third panel of Figure 1 shows that the corresponding VI vector field remains monotone, so simple fixed-point iterations can still converge to the global solution.

## 4 Statistical Guarantees of VI Estimators

Having established the formulation in (VI) and its motivation, we now turn to the theoretical analysis of the proposed estimator's statistical properties. This section derives finite-sample (Section 4.1) and

asymptotic (Section 4.2) guarantees for the empirical VI solution, defined as

$$\hat{\boldsymbol{\beta}}_N \in \mathrm{Sol}(V_N),$$

i.e., $V_N(\hat{\boldsymbol{\beta}}_N) = \frac{1}{N}\sum_{i=1}^N \big(g^{-1}(\tilde{\boldsymbol{x}}^{i\top}\hat{\boldsymbol{\beta}}_N) - y^i\big)\tilde{\boldsymbol{x}}^i = \mathbf{0}$. We begin with the following assumptions, which are used throughout the paper:

**Assumption 4.1.**

(i) The inverse link function $g^{-1}$ is $L$-Lipschitz continuous.

(ii) The true parameter $\boldsymbol{\beta}^\star$ and $(\boldsymbol{x}, y) \sim \mathbb{P}$ almost surely satisfy $|g^{-1}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}^\star) - y| \le R$.

(iii) There exists $M \ge 1$ such that $|x_j| \le M$ almost surely for all $j \in [d]$.

(iv) The population operator $V$ and the empirical operator $V_N$ satisfy the strong Minty condition with modulus $\mu > 0$ for any $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$:

$$\langle V(\boldsymbol{\beta}), \boldsymbol{\beta} - \boldsymbol{\beta}^\star\rangle \ge \mu\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|^2, \quad \langle V_N(\boldsymbol{\beta}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N\rangle \ge \mu\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N\|^2.$$

Consequently, the solution sets $\mathrm{Sol}(V)$ and $\mathrm{Sol}(V_N)$ are nonempty and singleton.

*Remark* 4.2 (Strong Minty condition). The strong Minty condition is a relatively mild regularity requirement compared to strong monotonicity. In Assumption 4.1, we adopt a slightly stronger version to ensure the uniqueness of both the true parameter $\boldsymbol{\beta}^\star$ and the sample-averaged solution $\hat{\boldsymbol{\beta}}_N$ for simplicity. Nevertheless, this assumption can be relaxed to the standard projection-based form in Definition C.1, by replacing $\boldsymbol{\beta}^\star$ and $\hat{\boldsymbol{\beta}}_N$ with $\mathrm{proj}_{\mathrm{Sol}(V)}(\boldsymbol{\beta})$ and $\mathrm{proj}_{\mathrm{Sol}(V_N)}(\boldsymbol{\beta})$, respectively. Intuitively, it requires that the vector field $V$ exhibits at least linear growth away from the solution, with a uniform positive coefficient that prevents flat or degenerate regions around the solution. Geometrically, in the one-dimensional case, this means that $V(\beta)$ has the same sign as $\beta - \beta^\star$ and satisfies $|V(\beta)| \ge \mu|\beta - \beta^\star|$, thereby providing a linear error bound around the solution. Further discussion and sufficient conditions are provided in Appendix C; see also related treatments in [10].

### 4.1 Finite-sample estimation error

We first derive the following estimation error bound for the variational inequality (VI).

**Theorem 4.3** (Estimation error)**.** *Suppose that Assumption 4.1 holds. Then for any $\epsilon \in (0,1)$, with probability at least $1-\epsilon$, we have the following estimation error*

$$\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star\| \le \frac{RM}{\mu}\sqrt{\frac{2(d+1)\ln(2(d+1)/\epsilon)}{N}}.$$

*Remark* 4.4. For the MLE, a similar finite-sample error bound can be derived as $\tilde{\mathcal{O}}(\sqrt{d}/(\mu'\sqrt{N}))$, where $\mu' > 0$ denotes the curvature modulus of the composite mapping $\ell \circ g^{-1}$ that satisfies the restricted secant inequality (RSI), i.e., for the maximum likelihood estimator $\bar{\boldsymbol{\beta}}_N$ and any $\boldsymbol{\beta}$

$$\langle \nabla\mathcal{L}_N(\boldsymbol{\beta}), \boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_N\rangle \ge \mu'\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_N\|^2,$$

a condition stronger than the Polyak–Łojasiewicz inequality and weaker than the strong convexity; see [16]. Recall that the gradient of the empirical loss of (MLE) is given by

$$\nabla\mathcal{L}_N(\boldsymbol{\beta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}[(g^{-1})'(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta})\ell'(g^{-1}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}), y)\tilde{\boldsymbol{x}}].$$

If the RSI holds with constant $\mu'$, then $\mu'\|\bar{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star\|^2 \le \langle \nabla\mathcal{L}_N(\boldsymbol{\beta}^\star), \boldsymbol{\beta}^\star - \bar{\boldsymbol{\beta}}_N\rangle$, which implies the bound $\|\bar{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star\| \le \|\nabla\mathcal{L}_N(\boldsymbol{\beta}^\star)\|/\mu'$. Compared with the VI estimator in Theorem 4.3, where by Lemma C.4 the strong Minty condition reduces to a condition on $g^{-1}$ itself, the MLE requires the composite function $z \mapsto (g^{-1})'(z)\ell'(g^{-1}(z), y)$ to satisfy an analogous Minty-type condition. This composite structure makes the regularity requirement for MLE substantially more intricate.

### 4.2 Asymptotic normality of VI estimators

We now proceed to establish the asymptotic normality of the VI estimator. Our goal is to show that $\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star)$ converges in distribution to a normal random vector with mean zero and a sandwich-type covariance matrix. To handle potential nonsmoothness of the link function $g^{-1}$, we

employ tools from empirical process theory and a local linearization argument around the population solution $\boldsymbol{\beta}^\star$. Since $g^{-1}$ is globally Lipschitz by Assumption 4.1, Rademacher's theorem [24, Theorem 9.60] implies that it is differentiable almost everywhere on $\mathbb{R}$, and hence its nondifferentiability set

$$\mathcal{K} := \{t : (g^{-1})'(t) \text{ does not exist}\}$$

has Lebesgue measure zero. We further assume the following condition.

**Assumption 4.5.** The random variable $\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta}^\star \rangle$ does not fall on the nondifferentiability set $\mathcal{K}$ of $g^{-1}$, that is,

$$\mathbb{P}\big(\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta}^\star \rangle \in \mathcal{K}\big) = 0.$$

This assumption is not restrictive in practice, since $\mathcal{K}$ has Lebesgue measure zero and thus the event $\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta}^\star \rangle \in \mathcal{K}$ occurs with probability zero, for instance, whenever $\boldsymbol{x}$ has a density and $\boldsymbol{\beta}^\star \neq \mathbf{0}$. With this mild condition, the population VI operator $V$ is guaranteed to be differentiable at the true parameter $\boldsymbol{\beta}^\star$.

**Lemma 4.6.** *Suppose that Assumptions 4.1 and 4.5 hold. Then the function $V$ is Fréchet differentiable at $\boldsymbol{\beta}^\star$ with Jacobian*

$$\nabla V(\boldsymbol{\beta}^\star) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[(g^{-1})'(\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta}^\star \rangle)\, \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right].$$

Owing to the averaging-induced smoothing effect, we obtain the asymptotic normality result.

**Theorem 4.7** (Asymptotic normality of the VI estimator)**.** *Suppose Assumptions 4.1 and 4.5 hold, and in addition the covariance matrix*

$$\Gamma := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}[V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)^\top] = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}[(y - g^{-1}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}^\star))^2\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top]$$

*is finite and positive definite. Then $\hat{\boldsymbol{\beta}}_N$ is $\sqrt{N}$-consistent and*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \nabla V(\boldsymbol{\beta}^\star)^{-1}\Gamma\nabla V(\boldsymbol{\beta}^\star)^{-\top}).$$

The resulting asymptotic covariance exhibits a sandwich structure analogous to that of the MLE. Under correct likelihood specification and standard regularity conditions, MLE is asymptotically efficient, whereas the VI estimator may incur an efficiency loss. We provide a detailed efficiency comparison in Appendix E. This potential loss can be viewed as a trade-off for superior computational performance with global convergence guarantees.

## 5 Convergence of VI-Based Algorithms

This section investigates the computational efficiency of solving the empirical VI problem using iterative algorithms. We begin by analyzing the convergence behavior of a fixed-point iterative method under a fixed sample size $N$. We then extend the analysis to the stochastic approximation setting, where the estimation is performed using a streaming data scheme.

For the following deterministic fixed-point iterative scheme under a fixed sample size, we establish its linear convergence rate under the strong Minty condition.

**Algorithm 1** Fixed-Point Iterative Method

**Input:** initialization $\boldsymbol{\beta}^0$, step size $\eta > 0$, data distribution $(\boldsymbol{x}, y) \sim \mathbb{P}_N$
**for** $t = 0$ **to** $T - 1$ **do**
  $\boldsymbol{\beta}^{t+1} \leftarrow \boldsymbol{\beta}^t - \eta\, V_N(\boldsymbol{\beta}^t)$
**end for**

**Theorem 5.1** (Linear convergence of fixed-point method)**.** *Suppose that Assumption 4.1 holds and set the step size $\eta = \frac{\mu}{L^2(1+dM^2)^2}$, where $\mu > 0$ is the strong Minty modulus in Assumption 4.1. Then the sequence of estimates $\boldsymbol{\beta}^t$ given by the Algorithm 1 satisfies for every $t = 0, 1, 2, \ldots,$*

$$\|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N\|^2 \leq \left(1 - \frac{\mu^2}{L^2(1+dM^2)^2}\right)^t \cdot \|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_N\|^2.$$

*Furthermore, for any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$, we have the following estimation error*

$$\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\| \leq \left(1 - \frac{\mu^2}{L^2(1+dM^2)^2}\right)^{t/2} \cdot \|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_N\| + \frac{RM}{\mu}\sqrt{\frac{2(d+1)\ln(2(d+1)/\epsilon)}{N}}.$$

Next, we consider the stochastic approximation setting, in which the estimator is updated from streaming data.

**Algorithm 2** Stochastic Approximation

**Input:** initialization $\boldsymbol{\beta}^0$, step sizes $\{\eta^t\}_{t=0}^{T-1}$ with $\eta^t > 0$, streaming data $(\boldsymbol{x}^t, y^t) \sim \mathbb{P}$
**for** $t = 0$ **to** $T-1$ **do**
  $\boldsymbol{\beta}^{t+1} \leftarrow \boldsymbol{\beta}^t - \eta^t V_{(\boldsymbol{x}^t, y^t)}(\boldsymbol{\beta}^t)$
**end for**

**Theorem 5.2** (Sublinear estimation error of stochastic approximation)**.** *Suppose that Assumption 4.1 holds. Then for any deterministic initialization $\boldsymbol{\beta}^0$, the sequence $\{\boldsymbol{\beta}^t\}_{t\geq 0}$ generated by Algorithm 2 from i.i.d. streaming data with step size $\eta^t = \frac{1}{\mu(t+1)}$, where $\mu > 0$ is the strong Minty modulus in Assumption 4.1, satisfies*

$$\mathbb{E}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2 \leq \frac{c_0(d+1)R^2M^2}{\mu^2(t+1)}, \qquad t = 0, 1, 2, \ldots,$$

*for some constant $c_0 > 0$ independent of $t$.*

## 6 Extension to Generalized Additive Models

Generalized additive models (GAMs) [8] extend GLMs by replacing the linear predictor with an additive (component-wise) structure, while retaining the exponential-family modeling of the response. A GAM links the conditional mean to the covariates via

$$\mathbb{E}[y \mid \boldsymbol{x}] = g^{-1}\left(\beta_0 + \sum_{j=1}^{d} f_j(x_j)\right),$$

where each additive component $f_j : \mathbb{R} \to \mathbb{R}$ describes the nonlinear effect of the $j$-th feature. To remove the ambiguity between the intercept and the additive constants, we impose an identifying constraint on each component, such as $f_j(0) = 0$ or centering with respect to a fixed reference measure.

**Basis representation.** To render the estimation problem finite-dimensional, we approximate each component function $f_j$ using a truncated basis expansion. Let $\{\psi_k\}_{k=1}^K$ be a set of basis functions (e.g., B-splines or polynomials) satisfying the chosen identifying constraint. For example, the Legendre basis functions $\psi_k = P_k$, $k = 1, \ldots, K$, satisfy $\int_{-1}^{1} \psi_k(u)\, du = 0$. We parameterize

$$f_j(x_j) \approx \sum_{k=1}^{K} \beta_{jk}\psi_k(x_j), \quad j = 1, \ldots, d.$$

Define the stacked feature mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^{dK}$ by

$$\Phi(\boldsymbol{x}) := [\psi_1(x_1), \ldots, \psi_K(x_1),\ \ldots,\ \psi_1(x_d), \ldots, \psi_K(x_d)]^\top,$$

and $\tilde{\Phi}(\boldsymbol{x}) := [1; \Phi(\boldsymbol{x})] \in \mathbb{R}^{dK+1}$. Let $\boldsymbol{\beta} \in \mathbb{R}^{dK+1}$ denote the augmented parameter vector collecting the intercept and all basis coefficients, i.e., $\boldsymbol{\beta} = (\beta_0, \{\beta_{jk}\}_{j\in[d],k\in[K]})$. Under this formulation, the additive predictor becomes linear in the transformed feature space:

$$\beta_0 + \sum_{j=1}^{d}\sum_{k=1}^{K} \beta_{jk}\psi_k(x_j) = \langle \boldsymbol{\beta}, \tilde{\Phi}(\boldsymbol{x})\rangle,$$

and hence the basis-expanded GAM reduces to a GLM problem in $\mathbb{R}^{dK+1}$.

The choice of basis functions impacts the theoretical properties of the estimator. We require the basis functions to be orthogonal (or at least linearly independent) with respect to the data distribution. If the basis functions are linearly dependent with respect to the data distribution, the second-moment matrix $\mathbb{E}[\tilde{\Phi}(\boldsymbol{x})\tilde{\Phi}(\boldsymbol{x})^\top]$ is singular, $V$ vanishes on a whole line through $\boldsymbol{\beta}^\star$, and the strong Minty condition in Assumption 4.1(iv) fails for every $\mu > 0$, so the uniqueness and stability guarantees of the VI estimator are lost.

**VI estimator for GAM.** Following the framework established for GLMs, we define the VI estimator for GAMs by substituting the original covariates with the basis features. The empirical vector field $V_N : \mathbb{R}^{dK+1} \to \mathbb{R}^{dK+1}$ is

$$V_N(\boldsymbol{\beta}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}\left[\left(g^{-1}\big(\langle\boldsymbol{\beta}, \tilde{\Phi}(\boldsymbol{x})\rangle\big) - y\right)\cdot\tilde{\Phi}(\boldsymbol{x})\right].$$

The VI estimator $\hat{\boldsymbol{\beta}}_N$ is then defined as the solution of $V_N(\boldsymbol{\beta}) = \mathbf{0}$, which is unique under Assumption 4.1(iv). This formulation allows us to extend the theoretical guarantees of the VI estimator to the GAM setting. In particular, when the basis size $K$ is fixed, the results in Sections 4 and 5 carry over with $d+1$ replaced by $dK+1$ and with the constants $M$ and $\mu$ in Assumption 4.1 referring to the transformed features $\tilde{\Phi}(\boldsymbol{x})$.

# 7 Numerical Experiments

We empirically evaluate the VI estimator from two complementary perspectives. First, we use controlled synthetic Poisson GLMs to isolate the optimization phenomena predicted by our analysis, such as flatness, truncation, and non-convexity induced by non-canonical links. Second, we evaluate the method on a real count-prediction benchmark to examine whether these stability gains can translate into improved predictive performance in practical GLM/GAM settings. Additional experimental details and supplementary results are provided in Appendix G. Code and instructions for reproducing the experiments are available at `https://github.com/jonghyeoki/vi-glm-estimation`.

## 7.1 Synthetic Poisson GLMs

We begin with synthetic Poisson regression experiments under several link functions. For each experiment, the true parameter is fixed as $\boldsymbol{\beta}^\star = d^{-1/2}(1,\dots,1) \in \mathbb{R}^d$, and the covariates $\boldsymbol{x}^i \in \mathbb{R}^d$, $i = 1,\dots,N$, are drawn i.i.d. from $\mathcal{N}(0, I_d)$. Given $\boldsymbol{x}^i$, the response is generated according to $y^i \sim \mathrm{Poisson}(g^{-1}(\boldsymbol{\beta}^{\star\top}\boldsymbol{x}^i))$, with no intercept term. We consider the following inverse link functions: (i) the log link $g^{-1}(z) = e^z$; (ii) the softplus link $g^{-1}(z) = \log(1+e^z)$; (iii) the clipped exponential link $g^{-1}(z) = \min\{e^z, 2\}$; and (iv) a scaled Gaussian-mixture CDF link $g^{-1}(z) = 4.5\Phi\big(\frac{z+1.0}{0.7}\big) + 5.5\Phi\big(\frac{z-1.2}{0.5}\big)$, where $\Phi$ denotes the standard Gaussian CDF. The log link corresponds to the canonical Poisson model, for which the VI and MLE updates coincide. The softplus and clipped exponential links introduce bounded curvature and truncation in the mean function, respectively, while the Gaussian-mixture CDF provides a non-convex example designed to test robustness in more challenging settings.

For each link, we compute both estimators by first-order iterative updates

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \eta^k V_N(\boldsymbol{\beta}^k), \quad \boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \eta^k \nabla\mathcal{L}_N(\boldsymbol{\beta}^k),$$

using the same exponentially decaying step sizes. We vary the dimension $d \in \{10, 20, 50, 100\}$ and sample size $N \in \{100, 200, 500, 1000\}$.

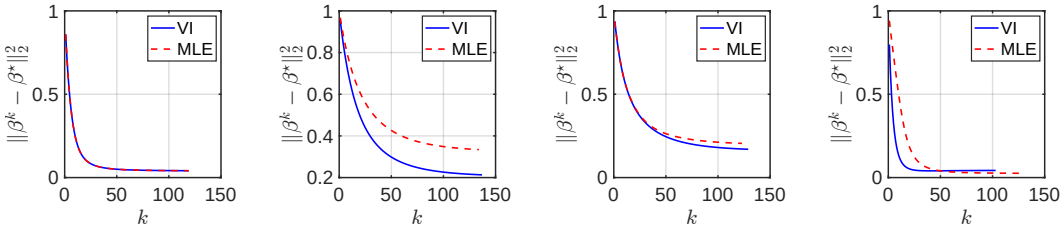
Table 2 reports the average squared error $\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^\star\|^2$ after $k \in \{20, 50, 100, 200\}$ iterations for the softplus link. The VI estimator consistently attains lower error than MLE across iteration budgets, often by a larger margin at the early stage. A more comprehensive set of results with other link functions and a sparse parameter setting can be found in Appendix G.1, where we observe that VI generally dominates MLE, except for the Gaussian-mixture CDF link in the "easy" regime of large sample size, low dimension, and large iterations.

Figure 2 and Figure 3 illustrate representative convergence behavior for $(d, N) = (20, 400)$. Figure 2 shows the sample convergence trajectories of the VI and MLE iterates for all four link functions. The trajectories for the log link coincide exactly, as expected from their equivalence as in Appendix B.1. For non-canonical links, the VI update exhibits noticeably faster convergence in the early iteration phase. In the clipped exponential case, the VI and MLE trajectories coincide during the initial iterations because the iterates have not yet reached the truncation boundary, and the model effectively behaves like the log link in that region. For the Gaussian-mixture CDF link, which induces a highly non-convex regime, VI avoids the oscillations and occasional divergence observed in MLE.

Figure 3 reports the average squared error against the iteration count $k$. The gap between VI and MLE is particularly pronounced for the softplus link, where VI converges significantly faster. For the

Table 2: Mean squared error of the VI estimator and MLE with softplus link across iterations $k$. For each $(k, d, N)$ combination, the smaller error between the two estimators is highlighted in bold. The values in parentheses are standard deviations across 1000 independent repetitions.

| $k$ | $d$ | $N = 100$ | | $N = 200$ | | $N = 500$ | | $N = 1000$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | VI | MLE | VI | MLE | VI | MLE | VI | MLE |
| 20 | 10 | **.63** (.08) | .71 (.06) | **.51** (.07) | .62 (.06) | **.34** (.06) | .47 (.05) | **.21** (.04) | .34 (.03) |
| | 20 | **.73** (.07) | .79 (.05) | **.63** (.06) | .71 (.05) | **.47** (.05) | .58 (.04) | **.33** (.04) | .46 (.03) |
| | 50 | **.82** (.05) | .87 (.03) | **.75** (.04) | .81 (.03) | **.63** (.04) | .71 (.03) | **.51** (.03) | .62 (.03) |
| | 100 | **.88** (.03) | .91 (.02) | **.83** (.03) | .87 (.02) | **.73** (.03) | .79 (.02) | **.63** (.03) | .71 (.02) |
| 50 | 10 | **.47** (.10) | .57 (.09) | **.32** (.08) | .44 (.07) | **.16** (.05) | .27 (.05) | **.08** (.03) | .16 (.03) |
| | 20 | **.59** (.09) | .67 (.07) | **.46** (.07) | .57 (.06) | **.28** (.05) | .40 (.05) | **.16** (.03) | .27 (.03) |
| | 50 | **.75** (.06) | .80 (.05) | **.64** (.06) | .71 (.05) | **.46** (.05) | .57 (.04) | **.32** (.04) | .44 (.03) |
| | 100 | **.83** (.05) | .86 (.04) | **.74** (.05) | .79 (.04) | **.59** (.04) | .68 (.03) | **.46** (.03) | .56 (.03) |
| 100 | 10 | **.40** (.11) | .50 (.10) | **.26** (.08) | .36 (.08) | **.12** (.04) | .20 (.04) | **.05** (.02) | .11 (.03) |
| | 20 | **.54** (.10) | .62 (.08) | **.40** (.08) | .50 (.07) | **.22** (.05) | .32 (.05) | **.12** (.03) | .20 (.03) |
| | 50 | **.71** (.07) | .76 (.06) | **.59** (.06) | .66 (.05) | **.40** (.05) | .50 (.04) | **.26** (.04) | .36 (.03) |
| | 100 | **.82** (.05) | .84 (.04) | **.72** (.05) | .76 (.04) | **.54** (.05) | .63 (.04) | **.40** (.04) | .50 (.03) |
| 200 | 10 | **.38** (.12) | .48 (.10) | **.24** (.08) | .34 (.08) | **.10** (.04) | .18 (.04) | **.05** (.02) | .09 (.02) |
| | 20 | **.53** (.11) | .61 (.09) | **.38** (.08) | .47 (.07) | **.20** (.04) | .30 (.04) | **.10** (.03) | .18 (.03) |
| | 50 | **.71** (.07) | .75 (.06) | **.57** (.07) | .64 (.06) | **.38** (.05) | .48 (.04) | **.24** (.04) | .34 (.03) |
| | 100 | **.82** (.06) | .84 (.04) | **.70** (.06) | .75 (.04) | **.52** (.05) | .61 (.04) | **.38** (.04) | .48 (.03) |




(a) log (b) softplus (c) clipped exp. (d) GMM CDF

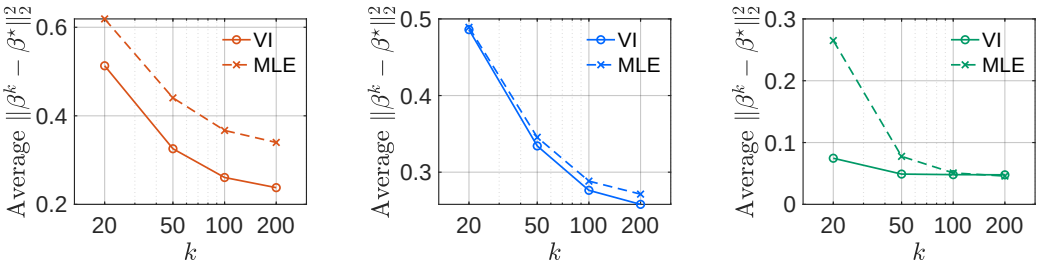
Figure 2: Convergence trajectories of VI and MLE for Poisson regression with different link functions ($d = 20, N = 400$).

clipped exponential link, the gap becomes more pronounced in the later iterations as the iterations approach the truncation boundary. For the Gaussian-mixture CDF link, the gap is largest during the early iterations and gradually narrows as both methods stabilize.

Overall, under the reported first-order iteration budgets, the results demonstrate that the proposed VI framework achieves comparable estimation accuracy to MLE under finite sample sizes, while often exhibiting faster convergence and greater numerical stability for non-canonical link functions. With canonical links, both methods coincide, but in more general GLMs, the VI formulation effectively mitigates the flat-landscape and conditioning issues that slow gradient-based likelihood estimation. Appendix G.5 also compares numerically converged softplus fits: at convergence, MLE attains slightly lower parameter error than VI, as expected from its asymptotic efficiency under correct specification (Appendix E).

### 7.2 Synthetic GAMs

We examine the reconstruction of a GAM using the VI and MLE estimators in a toy experiment. Consider a GAM of the form $\mathbb{E}[y \mid \boldsymbol{x}] = g^{-1}(\sum_{j=1}^{d} f_j(x_j))$. Each additive component $f_j$ is expanded using a truncated Legendre basis with the constant basis function $P_0$ omitted: $f_j(x_j) = \sum_{k=1}^{K} \beta_{jk} P_k(x_j)$, where $P_k$ denotes the $k$-th standard Legendre polynomial. With this basis expansion, the GAM reduces to a finite-dimensional GLM, to which we apply both the VI and MLE updates. Detailed experimental settings and results are deferred to Appendix G.3. Here we briefly summarize the main observations. For non-canonical links, VI typically exhibits more stable optimization behavior and slightly improved reconstruction of the underlying additive components under the tested first-order iteration budgets. These results demonstrate that the VI estimator naturally extends to GAMs and retains its key advantages beyond linear predictors.




(a) softplus (b) clipped exp. (c) GMM CDF

Figure 3: Average squared error for VI and MLE against iteration budgets for Poisson regression with different link functions under 500 replications ($d = 20, N = 400$).

### 7.3 Real data: Bike-sharing count prediction

We next evaluate whether the stability behavior observed in the synthetic experiments appears in a practical count-prediction task. We use the UCI hourly bike-sharing dataset [5], where the goal is to predict hourly rental counts from calendar and weather covariates. We use a chronological 3:1:1 train/validation/test split and compare GLM/GAM feature maps under the log, softplus, and clipped exponential links. For the GAM, continuous covariates are standardized using the training split and expanded using degree-3 Legendre bases, while categorical and time variables are one-hot encoded. For both VI and MLE, the initial step size is tuned by validation Poisson deviance.

The detailed results including runtime curves, calibration plots, and test metrics are provided in Appendix G.4. The main pattern is consistent with the synthetic results, that under the canonical log link, VI and MLE are essentially indistinguishable. Under the softplus link, VI obtains substantially better held-out performance and reaches a lower validation deviance at comparable runtime. The test-set calibration plot for the GAM design also shows a more stable predicted-versus-observed relationship for VI in the softplus case. The clipped exponential link gives a more mixed comparison that VI improves over MLE in the GLM design, but MLE is slightly better in the GAM design. These results again support that VI can provide substantial stability and predictive gains in harder non-canonical regimes under this fixed iteration budget. Appendix G.5 also compares converged softplus fits. Converged MLE substantially narrows the fixed-budget gap, but converged VI retains lower test deviance.

## 8 Discussion

This work shows that the behavior of the VI estimator is governed by the strong Minty modulus and the Lipschitz constant. These quantities determine convergence, stability, and statistical error bounds, suggesting a guideline for link-function design: with other quantities fixed, larger Minty moduli and smaller Lipschitz constants yield tighter guarantees. Compared to MLE, whose optimization depends on the composite loss $\ell \circ g^{-1}$, the VI framework offers greater flexibility in challenging regimes. In practice, VI can be particularly useful when likelihood objectives exhibit flat or non-smooth regions, while canonical models remain well served by MLE. The VI formulation decouples estimation from likelihood curvature, relying on the conditional mean rather than a fully specified likelihood. Nevertheless, under correct likelihood specification and standard regularity conditions, VI may incur an asymptotic efficiency loss relative to MLE (Appendix E). Future work includes extending the framework to regularized high-dimensional GLMs through variational inclusions and proximal updates, as well as analyzing penalized VI estimators under conditional-mean misspecification.

## Acknowledgments and Disclosure of Funding

This work is partially supported by NSF grants DMS-2220495, CNS-2220387, and DMS-2134037, and the Coca-Cola Foundation. The authors declare no competing interests.

## References

[1] Jean-Bernard Baillon and Georges Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel Journal of Mathematics*, 26:137–150, 1977.

[2] Yang Cao and Yao Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2016.

[3] Xiuyuan Cheng, Tingnan Gong, and Yao Xie. Point processes with event time uncertainty. *arXiv preprint arXiv:2411.02694*, 2024.

[4] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003.

[5] Hadi Fanaee-T. Bike sharing. UCI Machine Learning Repository, 2013. `https://doi.org/10.24432/C5W894`.

[6] Peter J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B*, 46(2): 149–170, 1984.

[7] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[8] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3): 297–310, 1986.

[9] Niao He, Zaid Harchaoui, Yichen Wang, and Le Song. Point process estimation with mirror prox algorithms. *Applied Mathematics & Optimization*, 82(3):919–947, 2020.

[10] Kevin Huang and Shuzhong Zhang. Beyond monotone variational inequalities: Solution methods and iteration complexities. *Pacific Journal of Optimization*, 20(3):403–428, 2024.

[11] Anatoli Juditsky and Arkadi Nemirovski. Signal recovery by stochastic optimization. *Automation and Remote Control*, 80(10):1878–1893, 2019.

[12] Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, and Yao Xie. Convex parameter recovery for interacting marked processes. *IEEE Journal on Selected Areas in Information Theory*, 1(3): 799–813, 2020.

[13] Anatoli Juditsky, Arkadi Nemirovski, Yao Xie, and Chen Xu. Generalized generalized linear models: Convex estimation and online bounds. *arXiv preprint arXiv:2304.13793*, 2023.

[14] Sham M. Kakade, Adam Tauman Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, volume 24, pages 927–935, 2011.

[15] Adam Tauman Kalai and Ravi Sastry. The Isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

[16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[17] Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Stochastic Modelling and Applied Probability*. Springer, New York, NY, 2nd edition, 2003.

[18] Chuanhai Liu. Robit regression: A simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 227–238. Wiley, 2004.

[19] Mengqi Lou, Kabir Verchand, Sara Fridovich-Keil, and Ashwin Pananjady. Accurate, provable, and fast polychromatic tomographic reconstruction: A variational inequality approach. *SIAM Journal on Imaging Sciences*, 19(1):446–479, 2026.

[20] Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

[21] James M. McFarland, Yuwei Cui, and Daniel A. Butts. Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Computational Biology*, 9(7):e1003143, 2013.

[22] John Ashworth Nelder and Robert W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

[23] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, New York, NY, 2013.

[24] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009.

[25] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[26] Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking reward modeling in preference-based large language model alignment. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[27] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, NY, 1996.

[28] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, UK, 2018.

[29] Jonathan Y. Zhou and Yao Xie. Nonlinear sequence embedding by monotone variational inequality. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

# A Notation

The notation in the paper is standard. For a positive integer $d$, we write $[d] := \{1, \ldots, d\}$. Let $\mathbb{R}^d$ denote the $d$-dimensional Euclidean space equipped with the inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle := \boldsymbol{u}^\top \boldsymbol{v}$ and the induced norm $\|\boldsymbol{u}\| := \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle}$. We write $\|\cdot\|_\infty$ for the infinity norm. For a closed convex set $\mathcal{B} \subseteq \mathbb{R}^d$, the Euclidean projection of a point $\boldsymbol{\beta}$ onto $\mathcal{B}$ is defined as $\mathrm{proj}_{\mathcal{B}}(\boldsymbol{\beta}) := \mathrm{argmin}_{\boldsymbol{\beta}' \in \mathcal{B}} \|\boldsymbol{\beta}' - \boldsymbol{\beta}\|$, and the corresponding distance is given by $\mathrm{dist}(\boldsymbol{\beta}, \mathcal{B}) := \|\boldsymbol{\beta} - \mathrm{proj}_{\mathcal{B}}(\boldsymbol{\beta})\|$. For a general vector field $F : \mathbb{R}^d \to \mathbb{R}^d$, its solution set is defined as

$$\mathrm{Sol}(F) := \left\{ \hat{\boldsymbol{\beta}} \in \mathbb{R}^d : \langle F(\hat{\boldsymbol{\beta}}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \rangle \geq 0,\ \forall \boldsymbol{\beta} \in \mathbb{R}^d \right\} = \left\{ \hat{\boldsymbol{\beta}} \in \mathbb{R}^d : F(\hat{\boldsymbol{\beta}}) = \mathbf{0} \right\}.$$

We use $\nabla F(\boldsymbol{\beta})$ to denote the Jacobian of $F$ at $\boldsymbol{\beta}$. We consider i.i.d. samples $\{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$ drawn from an unknown distribution $\mathbb{P}$, where each covariate $\boldsymbol{x}^i \in \mathbb{R}^d$ and response $y^i \in \mathbb{R}$. The empirical measure is denoted by $\mathbb{P}_N := \frac{1}{N} \sum_{i=1}^N \delta_{(\boldsymbol{x}^i, y^i)}$. Denote by $V_{(\boldsymbol{x}^i, y^i)}$ the per-sample VI operator and by $\mathcal{L}_{(\boldsymbol{x}^i, y^i)}$ the per-sample MLE negative log-likelihood function. We use $\xrightarrow{d}$ to denote convergence in distribution. The notation $\mathcal{O}(\cdot)$ denotes the standard asymptotic order, and $\tilde{\mathcal{O}}(\cdot)$ further hides logarithmic factors. The notation $\mathcal{O}_p(\cdot)$ denotes stochastic order in probability, that is, $X_N = \mathcal{O}_p(a_N)$ means that the sequence $\{X_N / a_N\}$ is bounded in probability. Likewise, the notations $o(\cdot)$ and $o_p(\cdot)$ denote deterministic and stochastic convergence to zero, respectively.

# B Detailed MLE and VI Optimization Landscape Derivations

In this appendix, we supplement the motivation provided in Section 3 with detailed mathematical derivations and a rigorous analysis of the optimization landscapes. First, we explicitly derive the gradients for several classical GLMs—including logistic, linear, exponential, and Poisson regression—to demonstrate the precise equivalence between MLE and VI formulation under canonical link functions (Appendix B.1). Second, we investigate the pathological behaviors of MLE in non-canonical settings. We provide an analytical calculation of the Hessian for the softplus link to quantify the ill-conditioning (flatness) of the MLE objective.

## B.1 Examples of vector fields for canonical MLE and VI

This section provides several representative examples of empirical vector fields arising from the MLE and the corresponding VI formulations. For each GLM, we derive the gradient of the empirical NLL function $\mathcal{L}_N(\boldsymbol{\beta})$ and verify that it coincides with the VI operator $V_N(\boldsymbol{\beta})$ under the canonical link function.

### B.1.1 Logistic regression

For the binary logistic regression, the empirical NLL is given by

$$\mathcal{L}_N(\boldsymbol{\beta}) = -\frac{1}{N} \sum_{i=1}^N \Big( y^i \log g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) + (1 - y^i) \log\big(1 - g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)\big) \Big),$$

with gradient given by

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^N \frac{(g^{-1})'(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)}{g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)\big(1 - g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)\big)} \cdot \Big( g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) - y^i \Big)\, \tilde{\boldsymbol{x}}^i.$$

For the logistic (sigmoid) link function $g^{-1}(z) = 1/(1 + e^{-z})$, one can show that $(g^{-1})'(z) = g^{-1}(z)(1 - g^{-1}(z))$, which simplifies the gradient to

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^N \big( g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) - y^i \big)\, \tilde{\boldsymbol{x}}^i,$$

and the last term above is identical to $V_N(\boldsymbol{\beta})$ in this case.

### B.1.2 Linear regression

For the Gaussian (linear) model with an identity link, the empirical NLL is

$$\mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{2N} \sum_{i=1}^{N} \left(g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) - y^i\right)^2,$$

whose gradient is

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} (g^{-1})'(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) \left(g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) - y^i\right) \tilde{\boldsymbol{x}}^i.$$

With the identity link $g^{-1}(z) = z$, we have $(g^{-1})'(z) = 1$, leading to

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) - y^i\right) \tilde{\boldsymbol{x}}^i = V_N(\boldsymbol{\beta}).$$

### B.1.3 Exponential regression

For the exponential model, the empirical NLL with $\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i > 0$ is

$$\mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(\log g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) + y^i / g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)\right),$$

and the corresponding gradient is

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} (g^{-1})'(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) \left(\frac{1}{g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)} - \frac{y^i}{g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)^2}\right) \tilde{\boldsymbol{x}}^i.$$

For the inverse link $g^{-1}(z) = z^{-1}$, we have $(g^{-1})'(z) = -z^{-2}$, yielding

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = -\frac{1}{N} \sum_{i=1}^{N} \left(g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) - y^i\right) \tilde{\boldsymbol{x}}^i = -V_N(\boldsymbol{\beta}).$$

The gradient $\nabla \mathcal{L}_N(\boldsymbol{\beta})$ differs from $V_N(\boldsymbol{\beta})$ only by a sign, so minimizing the NLL is equivalent to solving $V_N(\boldsymbol{\beta}) = \mathbf{0}$, yielding the same optimal solution.

### B.1.4 Poisson regression

For the Poisson model, the empirical NLL is

$$\mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(-y^i \log g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) + g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)\right).$$

The gradient is

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} (g^{-1})'(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) \left(-\frac{y^i}{g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i)} + 1\right) \tilde{\boldsymbol{x}}^i.$$

With the exponential link $g^{-1}(z) = e^z$, we have $(g^{-1})'(z) = g^{-1}(z)$, so

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(g^{-1}(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^i) - y^i\right) \tilde{\boldsymbol{x}}^i = V_N(\boldsymbol{\beta}).$$

### B.2 Flatness of MLE optimization landscape

We show that even when standard smooth link functions are applied to improve curvature, the resulting optimization landscape of the MLE can still exhibit considerably slower convergence than the VI formulation. This phenomenon is illustrated using the softplus link function, under which the MLE score and the VI operator define different estimating equations.

Recall the softplus inverse link function:

$$g^{-1}(z) = \log(1 + e^z), \qquad (g^{-1})'(z) = \sigma(z) := \frac{1}{1 + e^{-z}},$$

and

$$(g^{-1})''(z) = \sigma(z)(1 - \sigma(z)).$$

For a single sample $(\boldsymbol{x}, y)$, the Poisson negative log-likelihood is

$$\mathcal{L}_{(\boldsymbol{x},y)}(\boldsymbol{\beta}) = -y \log(g^{-1}(z)) + g^{-1}(z), \quad z = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta},$$

with gradient and Hessian given by, respectively,

$$\nabla \mathcal{L}_{(\boldsymbol{x},y)}(\boldsymbol{\beta}) = \sigma(z) \left(1 - \frac{y}{\log(1 + e^z)}\right) \tilde{\boldsymbol{x}},$$

and

$$\nabla^2 \mathcal{L}_{(\boldsymbol{x},y)}(\boldsymbol{\beta}) = \left\{ \sigma(z)(1 - \sigma(z)) \left(1 - \frac{y}{\log(1 + e^z)}\right) + \frac{y\sigma(z)^2}{\log(1 + e^z)^2} \right\} \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top.$$

On the other hand, the VI estimating operator is $V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}) = (g^{-1}(z) - y)\tilde{\boldsymbol{x}}$. At the population level, suppose that the conditional mean is correctly specified, and write $z^\star := \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^\star$, so that $\mathbb{E}[y \mid \boldsymbol{x}] = g^{-1}(z^\star)$. Evaluating at $\boldsymbol{\beta}^\star$ gives

$$\nabla V(\boldsymbol{\beta}^\star) = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}}} \left[ \sigma(z^\star) \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top \right].$$

Taking conditional expectations in the preceding Hessian formula and evaluating at $\boldsymbol{\beta}^\star$, we know that the first term vanishes, and hence

$$\nabla^2 \mathcal{L}(\boldsymbol{\beta}^\star) = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}}} \left[ \frac{\sigma(z^\star)^2}{g^{-1}(z^\star)} \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top \right].$$

Since $g^{-1}(z) \geq \sigma(z)$ for all $z$, we have $\sigma(z) \geq \sigma(z)^2 / g^{-1}(z)$ pointwise. Consequently,

$$\nabla V(\boldsymbol{\beta}^\star) - \nabla^2 \mathcal{L}(\boldsymbol{\beta}^\star) = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}}} \left[ \left( \sigma(z^\star) - \frac{\sigma(z^\star)^2}{g^{-1}(z^\star)} \right) \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top \right] \succeq \mathbf{0}.$$

Moreover, the corresponding scalar weights satisfy

$$\frac{\sigma(z)}{\sigma(z)^2 / g^{-1}(z)} = \frac{g^{-1}(z)}{\sigma(z)} = \frac{\log(1 + e^z)(1 + e^z)}{e^z} \longrightarrow +\infty \qquad \text{as } z \to +\infty.$$

In particular, $\sigma(z) \to 1$, whereas $\sigma(z)^2 / g^{-1}(z) \sim 1/z$. Thus, in the right tail, the likelihood curvature weight decays relative to the VI Jacobian weight. This comparison illustrates a stronger local response of the VI operator in large-predictor regimes.

## C Characterization of Minty Conditions

To start with, we introduce the following definition, which is widely used for solving variational inequalities, e.g., [10].

**Definition C.1** (Strong Minty condition)**.** The vector field $F : \mathbb{R}^d \to \mathbb{R}^d$ satisfies the strong Minty condition with constant $\mu \geq 0$ if $\mathrm{Sol}(F) \neq \emptyset$ and

$$\langle F(\boldsymbol{\beta}), \boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta}) \rangle \geq \mu \cdot \mathrm{dist}^2(\boldsymbol{\beta}, \mathrm{Sol}(F)), \quad \forall \boldsymbol{\beta} \in \mathbb{R}^d.$$

The strong Minty condition plays a central role in the VI estimator. In this section, we first establish sufficient conditions under which this property holds and illustrate them through representative examples. We then discuss guiding principles for checking link functions that satisfy the strong Minty condition in practice.

**Definition C.2.** Given constants $\rho, \mu_{\mathrm{EB}} > 0$, we introduce the following properties:

- $F$ is said to be $\rho$-weakly monotone if
$$\langle F(\boldsymbol{\beta}_2) - F(\boldsymbol{\beta}_1),\, \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \rangle \geq -\rho \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2, \qquad \forall\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d.$$
- $F$ is said to satisfy the error bound property with constant $\mu_{\mathrm{EB}} > 0$ if
$$\|F(\boldsymbol{\beta}) - F(\mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta}))\| \geq \mu_{\mathrm{EB}} \cdot \mathrm{dist}(\boldsymbol{\beta},\, \mathrm{Sol}(F)), \qquad \forall\, \boldsymbol{\beta} \in \mathbb{R}^d.$$

Weak monotonicity is a mild regularity condition that allows certain non-monotone vector fields while still ensuring local stability of the VI solution. In contrast, the error bound condition provides a geometric separation property: it quantifies how far a point $\boldsymbol{\beta}$ is from the solution set in terms of the magnitude of the operator, implying that the vector field $F(\boldsymbol{\beta})$ is uniformly separated from zero along the direction of its projection onto the solution set. Intuitively, this means that $F$ grows at least linearly with the distance to the solution. With these two conditions, we have the following implication.

**Proposition C.3.** *Suppose that the vector field $F = \nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is the gradient of a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, that $\mathrm{Sol}(F) \neq \emptyset$, and that $F$ satisfies the $L$-Lipschitz condition, i.e.,*
$$\|F(\boldsymbol{\beta}_2) - F(\boldsymbol{\beta}_1)\| \leq L\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|, \qquad \forall\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d.$$
*Additionally, we assume that $F$ satisfies the $\rho$-weak monotonicity condition in Definition C.2 with $0 < \rho < L$, and the error bound property holds with constant $\mu_{\mathrm{EB}} > 0$. If $\rho L < \mu_{\mathrm{EB}}^2$, then the strong Minty condition holds with $\mu = \frac{\mu_{\mathrm{EB}}^2 - \rho L}{L - \rho}$.*

*Proof.* Since $F$ satisfies the $\rho$-weak monotonicity and $L$-Lipschitz conditions, we know that $\boldsymbol{\beta} \mapsto F(\boldsymbol{\beta}) + \rho\boldsymbol{\beta}$ is monotone and $(L + \rho)$-Lipschitz. Moreover,
$$F(\boldsymbol{\beta}) + \rho\boldsymbol{\beta} = \nabla \left( f(\boldsymbol{\beta}) + \frac{\rho}{2}\|\boldsymbol{\beta}\|^2 \right),$$
and the function $\boldsymbol{\beta} \mapsto f(\boldsymbol{\beta}) + \frac{\rho}{2}\|\boldsymbol{\beta}\|^2$ is convex. Fix $\boldsymbol{\beta} \in \mathbb{R}^d$ and choose any Euclidean projection $\mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})$. By the Baillon–Haddad theorem [1, Corollaire 10], we have
$$\begin{aligned} &\left\langle F(\boldsymbol{\beta}) - F(\mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})) + \rho(\boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})), \boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta}) \right\rangle \\ \geq\, &\frac{1}{L + \rho} \left\| F(\boldsymbol{\beta}) - F(\mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})) + \rho(\boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})) \right\|^2, \end{aligned}$$
which implies that
$$\begin{aligned} &\frac{L - \rho}{L + \rho} \left\langle F(\boldsymbol{\beta}) - F(\mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})), \boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta}) \right\rangle \\ \geq\, &\frac{1}{L + \rho}\|F(\boldsymbol{\beta}) - F(\mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta}))\|^2 - \frac{\rho L}{L + \rho}\|\boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})\|^2. \end{aligned}$$
Combining this with the error bound property and using $L - \rho > 0$, we obtain
$$\begin{aligned} &\left\langle F(\boldsymbol{\beta}) - F(\mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})), \boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta}) \right\rangle \\ \geq\, &\frac{\mu_{\mathrm{EB}}^2}{L - \rho}\|\boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})\|^2 - \frac{\rho L}{L - \rho}\|\boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})\|^2 \\ =\, &\frac{\mu_{\mathrm{EB}}^2 - \rho L}{L - \rho}\|\boldsymbol{\beta} - \mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})\|^2. \end{aligned}$$
Since $F(\mathrm{proj}_{\mathrm{Sol}(F)}(\boldsymbol{\beta})) = \mathbf{0}$, this is the desired strong Minty inequality. The proof is complete. □

It is straightforward to verify that if $F$ is strongly monotone, then it implies the error bound property, which characterizes the local growth behavior around the solution set and implies the strong Minty condition. Moreover, we provide an example showing that $F$ need not be monotone for the strong Minty condition to hold.

*Example.* In many practical systems, such as those encountered in signal processing and communications, the ideal input-output response is approximately linear. However, due to hardware imperfections or nonlinear circuit effects, small nonlinear distortions may arise. A simple model capturing such behavior is given by

$$g^{-1}(z) = z + 2\sin(z)\cos(z).$$

The derivative of this function is

$$(g^{-1})'(z) = 1 + 2(\cos^2(z) - \sin^2(z)) = 3\cos^2(z) - \sin^2(z),$$

which is not always positive; hence, $g^{-1}$ is not a monotone function. Nevertheless, let $z^\star = 0$ denote the solution to $g^{-1}(z) = 0$. Then

$$\langle g^{-1}(z) - g^{-1}(z^\star),\, z - z^\star\rangle = g^{-1}(z)z \geq z^2 + 2z\sin(z)\cos(z) = z(z + \sin(2z)) \geq \frac{1}{2}z^2.$$

This inequality confirms that, despite the lack of monotonicity, the mapping $g^{-1}$ satisfies the strong Minty condition with modulus $\mu = 1/2$.

The following lemma shows that the vector field $V_N$ satisfies the strong Minty condition on a given parameter set whenever a weighted design matrix is positive definite, where the weights are the slopes of secants of the inverse link function $g^{-1}$. With a fixed sample size $N$, we collect the covariates into the data matrix $\boldsymbol{X}_N \in \mathbb{R}^{N\times d}$ defined as

$$\boldsymbol{X}_N := \begin{bmatrix} x_1^1 & \cdots & x_d^1 \\ \vdots & \ddots & \vdots \\ x_1^N & \cdots & x_d^N \end{bmatrix},$$

where $x_j^i$ denotes the $j$-th covariate of the $i$-th observation, and write $\tilde{\boldsymbol{X}}_N := [\mathbf{1}\ \boldsymbol{X}_N]$.

**Lemma C.4** (Sufficient condition for the strong Minty condition)**.** *Suppose that Assumption 4.1(i) holds, that $V_N(\hat{\boldsymbol{\beta}}_N) = \mathbf{0}$, and let $\mathcal{B} \subset \mathbb{R}^{d+1}$ be a set containing $\hat{\boldsymbol{\beta}}_N$. For $i \in [N]$, write $z_i := \tilde{\boldsymbol{x}}^{i\top}\hat{\boldsymbol{\beta}}_N$ and let*

$$\gamma_i := \inf\left\{ \frac{g^{-1}(z) - g^{-1}(z_i)}{z - z_i} :\ z = \tilde{\boldsymbol{x}}^{i\top}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathcal{B},\ z \neq z_i \right\}$$

*be the infimum of the slopes of the secants of $g^{-1}$ from $z_i$ to the predictor values attained on $\mathcal{B}$ (and $\gamma_i := 0$ if this set is empty). Then, for any $\boldsymbol{\beta} \in \mathcal{B}$,*

$$\langle V_N(\boldsymbol{\beta}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N\rangle \ \geq\ \lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^N \gamma_i\, \tilde{\boldsymbol{x}}^i\tilde{\boldsymbol{x}}^{i\top}\right)\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N\|^2.$$

*In particular, if $g^{-1}$ is $\mu_g$-strongly monotone with $\mu_g > 0$, then for $\mathcal{B} = \mathbb{R}^{d+1}$ every $\gamma_i \geq \mu_g$, and the vector field $V_N$ satisfies the strong Minty condition on $\mathbb{R}^{d+1}$ with modulus $\mu_g\sigma_N^2/N$, where $\sigma_N$ is the minimal singular value of $\tilde{\boldsymbol{X}}_N$.*

*Proof.* Fix $\boldsymbol{\beta} \in \mathcal{B}$. Recall the definition that $V_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^N (g^{-1}(\tilde{\boldsymbol{x}}^{i\top}\boldsymbol{\beta}) - y^i)\cdot\tilde{\boldsymbol{x}}^i$. Since $V_N(\hat{\boldsymbol{\beta}}_N) = \mathbf{0}$, we have

$$\begin{aligned}
\langle V_N(\boldsymbol{\beta}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N\rangle &= \frac{1}{N}\sum_{i=1}^N \left(g^{-1}(\tilde{\boldsymbol{x}}^{i\top}\boldsymbol{\beta}) - g^{-1}(\tilde{\boldsymbol{x}}^{i\top}\hat{\boldsymbol{\beta}}_N)\right)\cdot\tilde{\boldsymbol{x}}^{i\top}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N) \\
&\geq \frac{1}{N}\sum_{i=1}^N \gamma_i \left|\tilde{\boldsymbol{x}}^{i\top}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N)\right|^2 \\
&= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N)^\top\left(\frac{1}{N}\sum_{i=1}^N \gamma_i\, \tilde{\boldsymbol{x}}^i\tilde{\boldsymbol{x}}^{i\top}\right)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N) \\
&\geq \lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^N \gamma_i\, \tilde{\boldsymbol{x}}^i\tilde{\boldsymbol{x}}^{i\top}\right)\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N\|^2.
\end{aligned}$$

This establishes the first claim. It remains to consider the case where $g^{-1}$ is $\mu_g$-strongly monotone and $\mathcal{B} = \mathbb{R}^{d+1}$. In this case, the set defining $\gamma_i$ is nonempty since $\tilde{\boldsymbol{x}}^i \neq \mathbf{0}$, and every secant slope of $g^{-1}$ is at least $\mu_g$ by the strong monotonicity, so that $\gamma_i \geq \mu_g$ for any $i \in [N]$. Hence

$$\frac{1}{N}\sum_{i=1}^{N} \gamma_i \, \tilde{\boldsymbol{x}}^i \tilde{\boldsymbol{x}}^{i\top} \succeq \frac{\mu_g}{N} \tilde{\boldsymbol{X}}_N^\top \tilde{\boldsymbol{X}}_N,$$

and the minimal eigenvalue of the right-hand side is $\mu_g \sigma_N^2 / N$. Then the strong Minty condition on $\mathbb{R}^{d+1}$ follows from the first claim. The proof is complete. □

A population version holds under analogous conditions. If $V(\boldsymbol{\beta}^\star) = \mathbf{0}$ and $\mathbb{E}\|\tilde{\boldsymbol{x}}\|^2 < \infty$, then the same argument, with $\hat{\boldsymbol{\beta}}_N$ replaced by $\boldsymbol{\beta}^\star$ and the average replaced by the expectation, shows that for any set $\mathcal{B}$ containing $\boldsymbol{\beta}^\star$ and any $\boldsymbol{\beta} \in \mathcal{B}$,

$$\langle V(\boldsymbol{\beta}), \boldsymbol{\beta} - \boldsymbol{\beta}^\star \rangle \geq \lambda_{\min}(\mathbb{E}[\gamma(\boldsymbol{x}) \tilde{\boldsymbol{x}} \tilde{\boldsymbol{x}}^\top]) \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|^2,$$

where $\gamma(\boldsymbol{x})$ is defined in the same way as $\gamma_i$, with $\tilde{\boldsymbol{x}}$ in place of $\tilde{\boldsymbol{x}}^i$ and $\boldsymbol{\beta}^\star$ in place of $\hat{\boldsymbol{\beta}}_N$. The expectation is finite since $|\gamma(\boldsymbol{x})| \leq L$ by Assumption 4.1(i).

*Remark* C.5. Each $\gamma_i$ is a one-dimensional quantity and can be computed once $\mathcal{B}$ is fixed. For $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N\| \leq r\}$ with $r > 0$, the predictor values $\tilde{\boldsymbol{x}}^{i\top}\boldsymbol{\beta}$ range over $[z_i - r\|\tilde{\boldsymbol{x}}^i\|,\ z_i + r\|\tilde{\boldsymbol{x}}^i\|]$. If $(g^{-1})' \geq c$ almost everywhere on this interval, then $\gamma_i \geq c$, since each secant slope is an average of $(g^{-1})'$. For the softplus, Gaussian-mixture CDF, and arctangent links, $(g^{-1})'$ is positive and continuous, so $\gamma_i > 0$ for every $i$ on such a ball, and the modulus in Lemma C.4 is positive whenever $\tilde{\boldsymbol{X}}_N$ has full column rank. This is how Assumption 4.1(iv) can hold on a bounded set for links that are not strongly monotone. For a clipped link and the same ball, $\gamma_i = 0$ only for the observations whose fitted predictor $z_i$ is saturated, while an observation whose predictors enter the flat region for some $\boldsymbol{\beta} \in \mathcal{B}$ keeps $\gamma_i > 0$, since the secant from an unsaturated $z_i$ into the flat region still has positive slope. Hence the strong Minty condition can hold on $\mathcal{B}$ whenever the unsaturated observations span $\mathbb{R}^{d+1}$, although $g^{-1}$ is not strongly monotone on any interval that overlaps a flat region.

# D Proofs for Section 4

**Lemma D.1.** *Suppose that Assumptions 4.1* (i)–(iii) *hold. Then, for almost every* $(\boldsymbol{x}, y)$*, we have*

$$\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\|_\infty \leq RM, \qquad \|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\| \leq \sqrt{d+1}RM,$$

*and for any* $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{d+1}$*,*

$$\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}_2) - V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}_1)\| \leq (LdM^2 + L)\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|.$$

*Consequently, for any* $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$*,*

$$\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta})\|^2 \leq 2(d+1)R^2M^2 + 2L^2(1+dM^2)^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|^2.$$

*Proof.* First, from Assumptions 4.1 (ii)–(iii) it follows that, almost surely, $\|\boldsymbol{x}\|_\infty \leq M$, $\|\boldsymbol{x}\| \leq \sqrt{d}M$, and

$$\left| g^{-1}\left(\beta_0^\star + \sum_{j=1}^{d} \beta_j^\star x_j\right) - y \right| \leq R.$$

Then we have the upper bound (assuming $M \geq 1$) that

$$\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\|_\infty \leq \left| g^{-1}\left(\beta_0^\star + \sum_{j=1}^{d} \beta_j^\star x_j\right) - y \right| \|\tilde{\boldsymbol{x}}\|_\infty \leq RM,$$

$$\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\| \leq \left| g^{-1}\left(\beta_0^\star + \sum_{j=1}^{d} \beta_j^\star x_j\right) - y \right| \|\tilde{\boldsymbol{x}}\| \leq \sqrt{d+1}\, RM.$$

Next, for the Lipschitz constant of $V_{(\boldsymbol{x},y)}$, one has that

$$\begin{aligned}\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}') - V_{(\boldsymbol{x},y)}(\boldsymbol{\beta})\| &= \left\| \left( g^{-1}\left(\beta_0' + \sum_{j=1}^d \beta_j' x_j\right) - g^{-1}\left(\beta_0 + \sum_{j=1}^d \beta_j x_j\right)\right)\tilde{\boldsymbol{x}} \right\| \\ &\le L\|\tilde{\boldsymbol{x}}\| \cdot \left| \beta_0' - \beta_0 + \sum_{j=1}^d (\beta_j' - \beta_j) x_j \right| \\ &\le L\|\tilde{\boldsymbol{x}}\|^2 \|\boldsymbol{\beta}' - \boldsymbol{\beta}\| \\ &\le (LdM^2 + L)\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|.\end{aligned}$$

Finally, combining these bounds gives

$$\begin{aligned}\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta})\|^2 &\le 2\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\|^2 + 2\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}) - V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\|^2 \\ &\le 2(d+1)R^2M^2 + 2L^2(1+dM^2)^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|^2.\end{aligned}$$

The proof is complete. □

**Proof of Theorem 4.3**

First, from the strong Minty condition in Assumption 4.1 (iv), we know that

$$\mu\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}_N\|^2 \le \langle V_N(\boldsymbol{\beta}^\star), \boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}_N\rangle \le \|V_N(\boldsymbol{\beta}^\star)\|\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}_N\|,$$

which implies

$$\mu\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}_N\| \le \|V_N(\boldsymbol{\beta}^\star)\|. \tag{3}$$

Recall the definition that for any $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$

$$V_N(\boldsymbol{\beta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N}\left[\left(g^{-1}\left(\beta_0 + \sum_{j=1}^d \beta_j x_j\right) - y\right)\cdot\tilde{\boldsymbol{x}}\right] = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_N} V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}).$$

Then we know from Lemma D.1 that, almost surely,

$$\|V_{(\boldsymbol{x}^i,y^i)}(\boldsymbol{\beta}^\star)\|_\infty \le RM, \qquad i = 1,\dots,N.$$

As the samples are i.i.d. and $V(\boldsymbol{\beta}^\star) = \mathbf{0}$, by applying Hoeffding's inequality [28, Theorem 2.2.6] to the coordinates of $V_{(\boldsymbol{x}^i,y^i)}(\boldsymbol{\beta}^\star)$, it follows that for any $t > 0$ and $j \in \{0,\dots,d\}$,

$$\begin{aligned}\mathbb{P}\left(|V_N(\boldsymbol{\beta}^\star)_j| \ge t\right) &= \mathbb{P}\left(|V_N(\boldsymbol{\beta}^\star)_j - V(\boldsymbol{\beta}^\star)_j| \ge t\right) \\ &= \mathbb{P}\left(V_N(\boldsymbol{\beta}^\star)_j - V(\boldsymbol{\beta}^\star)_j \ge t\right) + \mathbb{P}\left(-V_N(\boldsymbol{\beta}^\star)_j + V(\boldsymbol{\beta}^\star)_j \ge t\right) \\ &\le 2\exp\left(-\frac{Nt^2}{2R^2M^2}\right).\end{aligned}$$

Using the union bound (i.e., Boole's inequality),

$$\begin{aligned}\mathbb{P}\left(\|V_N(\boldsymbol{\beta}^\star)\| \ge t\right) &\le \mathbb{P}\left(\sqrt{d+1}\|V_N(\boldsymbol{\beta}^\star)\|_\infty \ge t\right) \\ &\le \sum_{j=0}^d \mathbb{P}\left(|V_N(\boldsymbol{\beta}^\star)_j| \ge \frac{t}{\sqrt{d+1}}\right) \le 2(d+1)\exp\left(-\frac{Nt^2}{2(d+1)R^2M^2}\right).\end{aligned}$$

Equivalently, for any $\epsilon \in (0,1)$, set $t = RM\sqrt{2(d+1)\ln(2(d+1)/\epsilon)/N}$. Then with probability at least $1-\epsilon$,

$$\|V_N(\boldsymbol{\beta}^\star)\| \le RM\sqrt{\frac{2(d+1)\ln(2(d+1)/\epsilon)}{N}}.$$

Combining this with (3), we derive the desired results.

**Proof of Lemma 4.6**

Let $u := \langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta}^\star \rangle$ and $\delta_t := \langle \tilde{\boldsymbol{x}}, \boldsymbol{h}_t \rangle$ for any sequence $\boldsymbol{h}_t \to \boldsymbol{h}$ as $t \downarrow 0$. Then

$$\frac{V(\boldsymbol{\beta}^\star + t\boldsymbol{h}_t) - V(\boldsymbol{\beta}^\star)}{t} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[\frac{g^{-1}(u + t\delta_t) - g^{-1}(u)}{t}\,\tilde{\boldsymbol{x}}\right]. \tag{4}$$

Since $\mathbb{P}(u \in \mathcal{K}) = 0$ by Assumption 4.5, $g^{-1}$ is differentiable at $u$ almost surely, and therefore

$$g^{-1}(u + t\delta_t) = g^{-1}(u) + (g^{-1})'(u)\, t\delta_t + o(t|\delta_t|) \quad \text{a.s.}$$

Hence, pointwise almost surely,

$$\frac{g^{-1}(u + t\delta_t) - g^{-1}(u)}{t}\,\tilde{\boldsymbol{x}} \;\to\; (g^{-1})'(u)\,\langle \tilde{\boldsymbol{x}}, \boldsymbol{h}\rangle\,\tilde{\boldsymbol{x}}. \tag{5}$$

Moreover, by the global Lipschitz property of $g^{-1}$ with constant $L$,

$$\left\|\frac{g^{-1}(u + t\delta_t) - g^{-1}(u)}{t}\,\tilde{\boldsymbol{x}}\right\| \le L\,|\delta_t|\,\|\tilde{\boldsymbol{x}}\| \le L\,\|\tilde{\boldsymbol{x}}\|^2\|\boldsymbol{h}_t\|.$$

Under Assumption 4.1, we have $\|\tilde{\boldsymbol{x}}\|^2 \le 1 + dM^2$, so the right-hand side is bounded by $L(1 + dM^2)\|\boldsymbol{h}_t\|$, which is integrable and independent of $t$ since $\|\boldsymbol{h}_t\|$ is bounded. From (4) and (5) with the dominated convergence theorem, it follows that

$$\frac{V(\boldsymbol{\beta}^\star + t\boldsymbol{h}_t) - V(\boldsymbol{\beta}^\star)}{t} \;\to\; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[(g^{-1})'(\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta}^\star\rangle)\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\boldsymbol{h}.$$

Thus $V$ is Fréchet differentiable at $\boldsymbol{\beta}^\star$ with Jacobian $\nabla V(\boldsymbol{\beta}^\star)$ as claimed.

**Proof of Theorem 4.7**

By Assumption 4.1, $|g^{-1}(\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta}^\star\rangle) - y| \le R$, $\|\boldsymbol{x}\|^2 \le dM^2$, and $g^{-1}$ is Lipschitz. Hence,

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[\|V_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\|^2\right] = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[|g^{-1}(\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta}^\star\rangle) - y|^2\|\tilde{\boldsymbol{x}}\|^2\right] \le R^2(1 + dM^2) < \infty.$$

Therefore, when $(\boldsymbol{x}^i, y^i)_{i=1}^N$ are i.i.d. samples, the random vectors $V_{(\boldsymbol{x}^i,y^i)}(\boldsymbol{\beta}^\star)$ are i.i.d. with mean zero and covariance matrix $\Gamma$ finite and positive definite. Applying the multivariate central limit theorem yields

$$\sqrt{N}\big(V_N(\boldsymbol{\beta}^\star) - V(\boldsymbol{\beta}^\star)\big) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \Gamma). \tag{6}$$

Adding and subtracting $V(\hat{\boldsymbol{\beta}}_N)$ and $V(\boldsymbol{\beta}^\star)$ in $V_N(\hat{\boldsymbol{\beta}}_N)$ yields

$$\boldsymbol{0} = V_N(\hat{\boldsymbol{\beta}}_N) = V_N(\boldsymbol{\beta}^\star) + \underbrace{V(\hat{\boldsymbol{\beta}}_N) - V(\boldsymbol{\beta}^\star)}_{\text{population increment}} + \underbrace{(V_N(\hat{\boldsymbol{\beta}}_N) - V(\hat{\boldsymbol{\beta}}_N)) - (V_N(\boldsymbol{\beta}^\star) - V(\boldsymbol{\beta}^\star))}_{=:R_N}. \tag{7}$$

By Assumption 4.1(iv) and Lemma 4.6, for any $\boldsymbol{h} \in \mathbb{R}^{d+1}$,

$$\boldsymbol{h}^\top \nabla V(\boldsymbol{\beta}^\star)\boldsymbol{h} = \lim_{t\downarrow 0} \frac{\langle V(\boldsymbol{\beta}^\star + t\boldsymbol{h}), t\boldsymbol{h}\rangle}{t^2} \ge \mu\|\boldsymbol{h}\|^2,$$

so $\nabla V(\boldsymbol{\beta}^\star)$ is non-singular. By Theorem 4.3, $\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star\| = O_p(N^{-1/2})$. The Fréchet differentiability of $V$ at $\boldsymbol{\beta}^\star$ (Lemma 4.6) therefore gives the local linearization

$$V(\hat{\boldsymbol{\beta}}_N) - V(\boldsymbol{\beta}^\star) = \nabla V(\boldsymbol{\beta}^\star)(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star) + r_N, \quad \|r_N\| = o_p(\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star\|). \tag{8}$$

From Lemma D.1, the sample operator $V_{(\boldsymbol{x},y)}$ is $(LdM^2 + L)$-Lipschitz. Write $L_V := LdM^2 + L$ and note that

$$R_N = [\mathbb{E}_N - \mathbb{E}]\left(V_{(\cdot)}(\hat{\boldsymbol{\beta}}_N) - V_{(\cdot)}(\boldsymbol{\beta}^\star)\right),$$

where $\mathbb{E}_N$ denotes averaging over the sample. For a fixed $0 < r \le 1$, define the local class

$$\mathcal{F}_r := \left\{ f_{\boldsymbol{\beta}}(\cdot) := V_{(\cdot)}(\boldsymbol{\beta}) - V_{(\cdot)}(\boldsymbol{\beta}^\star) :\ \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\| \le r \right\}.$$

For every coordinate $j \in \{0, \dots, d\}$, let $\mathcal{F}_{r,j} := \{f_{\boldsymbol{\beta},j} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\| \le r\}$, where $f_{\boldsymbol{\beta},j}$ is the $j$-th coordinate of $f_{\boldsymbol{\beta}}$. Almost surely,

$$|f_{\boldsymbol{\beta},j}(\boldsymbol{x}, y)| \le L_V r, \qquad |f_{\boldsymbol{\beta},j}(\boldsymbol{x}, y) - f_{\boldsymbol{\beta}',j}(\boldsymbol{x}, y)| \le L_V \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|.$$

By van der Vaart and Wellner [27, Theorem 2.7.11], the bracketing numbers of $\mathcal{F}_{r,j}$ in $L_2(\mathbb{P})$ are bounded by the covering numbers of the parameter ball, which are polynomial in $1/\epsilon$. Consequently, each $\mathcal{F}_{r,j}$ is $\mathbb{P}$-Donsker by van der Vaart and Wellner [27, Theorem 2.5.6]. Moreover, $f_{\boldsymbol{\beta}^\star} = \mathbf{0}$ and $(\mathbb{E}\|f_{\boldsymbol{\beta}}(\boldsymbol{x}, y)\|^2)^{1/2} \le L_V \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|$. The asymptotic equicontinuity of the coordinate empirical processes therefore implies that, for every $\epsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{N \to \infty} \mathbb{P}\left( \sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\| \le \delta} \left\| \sqrt{N}[\mathbb{E}_N - \mathbb{E}] f_{\boldsymbol{\beta}} \right\| > \epsilon \right) = 0, \tag{9}$$

where $0 < \delta \le r$. Since $\hat{\boldsymbol{\beta}}_N$ is consistent, for every such $\delta$ we have

$$\mathbb{P}(\sqrt{N}\|R_N\| > \epsilon) \le \mathbb{P}(\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star\| > \delta) + \mathbb{P}\left( \sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\| \le \delta} \left\| \sqrt{N}[\mathbb{E}_N - \mathbb{E}] f_{\boldsymbol{\beta}} \right\| > \epsilon \right).$$

Taking the limit superior as $N \to \infty$ and then letting $\delta \downarrow 0$, we conclude from (9) that $\sqrt{N}\|R_N\| = o_p(1)$, i.e., $\|R_N\| = o_p(N^{-1/2})$. Substituting (8) into (7) gives

$$\mathbf{0} = V_N(\boldsymbol{\beta}^\star) + \nabla V(\boldsymbol{\beta}^\star)(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star) + r_N + R_N.$$

Rearranging terms yields

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star) = -\nabla V(\boldsymbol{\beta}^\star)^{-1}\sqrt{N}\big(V_N(\boldsymbol{\beta}^\star) - V(\boldsymbol{\beta}^\star)\big) - \nabla V(\boldsymbol{\beta}^\star)^{-1}\sqrt{N}(r_N + R_N).$$

The first term converges in distribution to $\mathcal{N}(\mathbf{0}, \nabla V(\boldsymbol{\beta}^\star)^{-1}\Gamma\nabla V(\boldsymbol{\beta}^\star)^{-\top})$ by (6). On the other hand, $\sqrt{N}\|r_N\| = o_p(1)$ by (8) and $\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star\| = O_p(N^{-1/2})$, while $\sqrt{N}\|R_N\| = o_p(1)$ by the preceding argument. Thus, the second term vanishes in probability. Applying Slutsky's theorem gives

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \nabla V(\boldsymbol{\beta}^\star)^{-1}\Gamma\nabla V(\boldsymbol{\beta}^\star)^{-\top}).$$

The proof is complete.

# E Comparison of Asymptotic Efficiency between VI and MLE

A classical result in asymptotic statistics states that the MLE is asymptotically efficient, namely, it is consistent, asymptotically unbiased, and attains the Cramér–Rao lower bound (CRB) under correct model specification. The asymptotic covariance of the proposed VI estimator,

$$\Sigma_{\mathrm{VI}} = \nabla V(\boldsymbol{\beta}^\star)^{-1}\Gamma\nabla V(\boldsymbol{\beta}^\star)^{-\top},$$

shares the same "sandwich" structure as that of the MLE,

$$\Sigma_{\mathrm{MLE}} = \big(\nabla^2\mathcal{L}(\boldsymbol{\beta}^\star)\big)^{-1}\,\mathrm{Cov}(\nabla\mathcal{L}_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star))\big(\nabla^2\mathcal{L}(\boldsymbol{\beta}^\star)\big)^{-\top},$$

where the expected Hessian and the score covariance are given by

$$\nabla^2\mathcal{L}(\boldsymbol{\beta}^\star) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[\frac{(g^{-1})'(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}^\star)^2}{\mathrm{Var}(y \mid \tilde{\boldsymbol{x}})}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right],$$

and

$$\begin{aligned}\mathrm{Cov}(\nabla\mathcal{L}_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)) &:= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\big[\nabla\mathcal{L}_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\nabla\mathcal{L}_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)^\top\big] \\ &= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[\frac{(y - g^{-1}(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}^\star))^2(g^{-1})'(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}^\star)^2}{\mathrm{Var}(y \mid \tilde{\boldsymbol{x}})^2}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right].\end{aligned}$$

For correctly specified models, the information identity

$$\mathrm{Cov}\big(\nabla\mathcal{L}_{(\boldsymbol{x},y)}(\boldsymbol{\beta}^\star)\big) = \nabla^2\mathcal{L}(\boldsymbol{\beta}^\star) = I(\boldsymbol{\beta}^\star)$$

holds, and the MLE achieves the Fisher information bound $I(\boldsymbol{\beta}^\star)^{-1}$, attaining the CRB. In the following, we show that the MLE can indeed be asymptotically more efficient than the VI estimator under correct model specification. Their asymptotic covariances coincide only when the link function is canonical. This observation does not contradict our earlier discussion, as the VI estimator may still outperform the MLE in the presence of model or variance misspecification.

For correctly specified models with general link functions, the VI estimator generally does not satisfy the corresponding identity $\Gamma = \nabla V(\boldsymbol{\beta}^\star)$, since the residual covariance and the Jacobian $\nabla V(\boldsymbol{\beta}^\star)$ coincide only for canonical links. Let

$$\Sigma_{\mathrm{MLE}}^{-1} = \mathbb{E}\left[\frac{(g^{-1})'(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^\star)^2}{\mathrm{Var}(y \mid \tilde{\boldsymbol{x}})}\, \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right],$$

$$\boldsymbol{r}_1 = \sqrt{\mathrm{Var}(y \mid \tilde{\boldsymbol{x}})}\, \tilde{\boldsymbol{x}}, \quad \boldsymbol{r}_2 = \frac{(g^{-1})'(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^\star)}{\sqrt{\mathrm{Var}(y \mid \tilde{\boldsymbol{x}})}}\, \tilde{\boldsymbol{x}}.$$

Then $\mathbb{E}[\boldsymbol{r}_1 \boldsymbol{r}_1^\top] = \Gamma$, $\mathbb{E}[\boldsymbol{r}_2 \boldsymbol{r}_2^\top] = \Sigma_{\mathrm{MLE}}^{-1}$, $\mathbb{E}[\boldsymbol{r}_1 \boldsymbol{r}_2^\top] = \nabla V(\boldsymbol{\beta}^\star)$, and hence the block moment matrix

$$\begin{bmatrix} \Gamma & \nabla V(\boldsymbol{\beta}^\star) \\ \nabla V(\boldsymbol{\beta}^\star) & \Sigma_{\mathrm{MLE}}^{-1} \end{bmatrix} \succeq \mathbf{0}.$$

By the Schur complement, this implies the MLE is never less efficient than the VI estimator, i.e.,

$$\Sigma_{\mathrm{MLE}} \preceq \Sigma_{\mathrm{VI}}.$$

The equality holds if and only if $\boldsymbol{r}_2 = A\boldsymbol{r}_1$ almost surely for some fixed matrix $A$. In particular, it holds when, for a constant $c > 0$,

$$\mathbb{E}\big[(y - g^{-1}(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^\star))^2 \mid \tilde{\boldsymbol{x}}\big] = c \cdot (g^{-1})'(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^\star) \quad \text{a.s. in } \tilde{\boldsymbol{x}},$$

which is the case for canonical links (up to a constant scale), and for nondegenerate designs this condition is also necessary. Hence, under correct specification, the MLE dominates the VI estimator in efficiency, with equality for canonical links. As a result, the VI estimator may exhibit a loss of statistical efficiency relative to the MLE, for which the MLE is asymptotically optimal in the sense of attaining the CRB. Under misspecification, the comparison changes. If the conditional mean (1) is correct but the variance or the full conditional distribution is not, both estimators remain consistent for $\boldsymbol{\beta}^\star$, and neither has uniformly smaller asymptotic variance. If the conditional mean or the link is itself misspecified, the VI estimator converges to the solution of the moment equation $V(\boldsymbol{\beta}) = \mathbf{0}$, whereas the MLE converges to the Kullback–Leibler projection induced by the working likelihood.

# F Proofs for Section 5

**Proof of Theorem 5.1**

From the update rule of the fixed-point iterative method, the iterates satisfy

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \eta\, V_N(\boldsymbol{\beta}^t), \qquad \hat{\boldsymbol{\beta}}_N = \hat{\boldsymbol{\beta}}_N - \eta\, V_N(\hat{\boldsymbol{\beta}}_N),$$

where $\hat{\boldsymbol{\beta}}_N$ is a fixed point of the map $\boldsymbol{\beta} \mapsto \boldsymbol{\beta} - \eta V_N(\boldsymbol{\beta})$ since $V_N(\hat{\boldsymbol{\beta}}_N) = \mathbf{0}$. It then follows that

$$\|\boldsymbol{\beta}^{t+1} - \hat{\boldsymbol{\beta}}_N\|^2 = \|\boldsymbol{\beta}^t - \eta\, V_N(\boldsymbol{\beta}^t) - \hat{\boldsymbol{\beta}}_N + \eta\, V_N(\hat{\boldsymbol{\beta}}_N)\|^2.$$

Expanding the square gives

$$\|\boldsymbol{\beta}^{t+1} - \hat{\boldsymbol{\beta}}_N\|^2 = \|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N\|^2 - 2\eta\big\langle \boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N,\, V_N(\boldsymbol{\beta}^t) - V_N(\hat{\boldsymbol{\beta}}_N)\big\rangle + \eta^2 \|V_N(\boldsymbol{\beta}^t) - V_N(\hat{\boldsymbol{\beta}}_N)\|^2.$$

By the strong Minty condition in Assumption 4.1 and the Lipschitz continuity in Lemma D.1, we have

$$\langle V_N(\boldsymbol{\beta}^t) - V_N(\hat{\boldsymbol{\beta}}_N), \boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N\rangle \geq \mu \|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N\|^2, \quad \|V_N(\boldsymbol{\beta}^t) - V_N(\hat{\boldsymbol{\beta}}_N)\| \leq (LdM^2 + L)\|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N\|.$$

Substituting these bounds gives

$$\|\boldsymbol{\beta}^{t+1} - \hat{\boldsymbol{\beta}}_N\|^2 \leq \big(1 - 2\eta\mu + \eta^2 L^2 (1 + dM^2)^2\big)\|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N\|^2.$$

Choosing $\eta = \mu/(L^2(1+dM^2)^2)$ yields

$$\|\boldsymbol{\beta}^{t+1} - \hat{\boldsymbol{\beta}}_N\|^2 \leq \left(1 - \frac{\mu^2}{L^2(1+dM^2)^2}\right) \|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N\|^2.$$

By applying the above bound recursively, we obtain the claimed linear convergence rate of Algorithm 1. Combining this result with Theorem 4.3 further yields with probability at least $1-\epsilon$,

$$\begin{aligned}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\| &\leq \|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}_N\| + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^\star\| \\ &\leq \left(1 - \frac{\mu^2}{L^2(1+dM^2)^2}\right)^{t/2} \|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_N\| + \frac{RM}{\mu}\sqrt{\frac{2(d+1)\ln(2(d+1)/\epsilon)}{N}}.\end{aligned}$$

The proof is complete.

**Proof of Theorem 5.2**

Since $\boldsymbol{\beta}^0$ is deterministic, the growth bound in Lemma D.1 implies by induction that all iterates have finite second moments. From the update rule, we know that

$$\begin{aligned}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^\star\|^2 &= \|\boldsymbol{\beta}^t - \eta^t V_{(\boldsymbol{x}^t, y^t)}(\boldsymbol{\beta}^t) - \boldsymbol{\beta}^\star\|^2 \\ &= \|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2 - 2\eta^t \langle V_{(\boldsymbol{x}^t, y^t)}(\boldsymbol{\beta}^t), \boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\rangle + (\eta^t)^2 \|V_{(\boldsymbol{x}^t, y^t)}(\boldsymbol{\beta}^t)\|^2.\end{aligned}$$

Taking expectations and using the law of total expectation yields

$$\frac{1}{2}\mathbb{E}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^\star\|^2 \leq \frac{1}{2}\mathbb{E}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2 - \eta^t \mathbb{E}\left[\langle V_{(\boldsymbol{x}^t, y^t)}(\boldsymbol{\beta}^t), \boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\rangle\right] + \frac{1}{2}(\eta^t)^2 \mathbb{E}\|V_{(\boldsymbol{x}^t, y^t)}(\boldsymbol{\beta}^t)\|^2. \tag{10}$$

Since $(\boldsymbol{x}^t, y^t)$ is independent of $\boldsymbol{\beta}^t$, the strong Minty condition in Assumption 4.1 gives

$$\mathbb{E}\left[\langle V_{(\boldsymbol{x}^t, y^t)}(\boldsymbol{\beta}^t), \boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\rangle\right] = \mathbb{E}\left[\langle V(\boldsymbol{\beta}^t), \boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\rangle\right] \geq \mu \mathbb{E}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2.$$

In addition, Lemma D.1 implies that

$$\mathbb{E}\|V_{(\boldsymbol{x}^t, y^t)}(\boldsymbol{\beta}^t)\|^2 \leq 2(d+1)R^2M^2 + 2L^2(1+dM^2)^2 \mathbb{E}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2.$$

Substituting these bounds into (10) yields

$$\frac{1}{2}\mathbb{E}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^\star\|^2 \leq \frac{1}{2}\left(1 - 2\mu\eta^t + 2(\eta^t)^2 L^2(1+dM^2)^2\right)\mathbb{E}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2 + (\eta^t)^2(d+1)R^2M^2.$$

Let $\eta^t = 1/[\mu(t+1)]$. Then

$$\frac{1}{2}\mathbb{E}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^\star\|^2 \leq \frac{1}{2}\left(1 - \frac{2}{t+1} + \frac{2L^2(1+dM^2)^2}{\mu^2(t+1)^2}\right)\mathbb{E}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2 + \frac{(d+1)R^2M^2}{\mu^2(t+1)^2}.$$

Choose an integer

$$t_0 \geq \max\left\{1, \frac{4L^2(1+dM^2)^2}{\mu^2}\right\}.$$

For every $t \geq t_0$, we have

$$\frac{2L^2(1+dM^2)^2}{\mu^2(t+1)^2} \leq \frac{1}{2(t+1)},$$

and hence

$$\frac{1}{2}\mathbb{E}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^\star\|^2 \leq \frac{1}{2}\left(1 - \frac{3}{2(t+1)}\right)\mathbb{E}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2 + \frac{(d+1)R^2M^2}{\mu^2(t+1)^2}.$$

Since the first $t_0+1$ iterates have finite second moments, we can choose $c_0 \geq 4$ sufficiently large such that

$$\frac{1}{2}\mathbb{E}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|^2 \leq \frac{c_0(d+1)R^2M^2}{2\mu^2(t+1)}, \qquad t = 0, 1, \ldots, t_0.$$

The preceding recursion shows that $c_0$ can be chosen in terms of the normalized initial error $\mu^2\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|^2/[(d+1)R^2M^2]$ and $L(1+dM^2)/\mu$.

We now prove by induction that this bound holds for all $t \geq 0$. The cases $0 \leq t \leq t_0$ hold by the choice of $c_0$. For the induction step, suppose that the bound holds at some $t \geq t_0$. Since $1 - 3/[2(t+1)] \geq 0$, we obtain

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^\star\|^2 &\leq \frac{c_0(d+1)R^2M^2}{2\mu^2(t+1)}\left(1 - \frac{3}{2(t+1)}\right) + \frac{(d+1)R^2M^2}{\mu^2(t+1)^2} \\
&= \frac{c_0(d+1)R^2M^2}{2\mu^2}\left(\frac{1}{t+1} - \frac{3/2 - 2/c_0}{(t+1)^2}\right) \\
&\leq \frac{c_0(d+1)R^2M^2}{2\mu^2}\left(\frac{1}{t+1} - \frac{1}{(t+1)^2}\right) \\
&\leq \frac{c_0(d+1)R^2M^2}{2\mu^2(t+2)},
\end{aligned}$$

where the second inequality uses $c_0 \geq 4$. This completes the induction and thus the proof.

# G Additional Experimental Details and Results

## G.1 Additional synthetic GLM results

We provide additional results supplementing the synthetic Poisson GLM experiment in Section 7.1. Tables 3 and 4 present additional numerical results for the finite-sample performance of VI and MLE under different parameter structures. Table 3 corresponds to the dense parameter setting $\boldsymbol{\beta}^\star = d^{-1/2}(1, \ldots, 1) \in \mathbb{R}^d$ as in Section 7. The results for the softplus link were already discussed in Table 2, so here we report the outcomes for the log (sanity check), clipped exponential, and Gaussian-mixture CDF links. Table 4 corresponds to the sparse parameter setting, where $\boldsymbol{\beta}^\star = (2/\sqrt{5}, 1/\sqrt{5}, 0, \ldots, 0) \in \mathbb{R}^d$, and summarizes the results for the softplus, clipped exponential, and Gaussian-mixture CDF links (the log link is omitted).

Overall, the results under the sparse parameter setting are qualitatively similar to those under the dense parameter setting. In both cases, the VI estimator consistently outperforms MLE across most configurations, with particularly large improvements for the softplus link.

## G.2 MLE stalling under saturated initialization

The clipped-exponential link provides a simple example where likelihood-based optimization can become completely flat. In the main synthetic experiments of Section 7.1, the default initialization does not necessarily place the iterates inside the saturated region, and therefore the clipped-exponential experiment does not always show complete stalling at early iterations. To directly isolate this failure mode, we consider an additional stress test for the Poisson model with

$$g^{-1}(z) = \min\{\exp(z), C\}, \qquad C = 2.$$

We use $d = 20$ non-intercept features and $N = 400$ samples. The true parameter has zero intercept and dense unit-norm slope,

$$\boldsymbol{\beta}^\star = (0, d^{-1/2}, \ldots, d^{-1/2}) \in \mathbb{R}^{d+1}.$$

The covariates are drawn as $\boldsymbol{x}^i \sim \mathcal{N}(0, I_d)$, and we augment them with an intercept coordinate. This intercept coordinate is used only to construct a saturated initialization. Specifically, both VI and MLE are initialized at

$$\boldsymbol{\beta}^0 = (2.0,\ 0.25\,\boldsymbol{\beta}^\star_{2:d+1}),$$

which pushes the initial predictors $z_i^0 = \langle \boldsymbol{\beta}^0, \tilde{\boldsymbol{x}}^i \rangle$ beyond the clipping threshold $\log C$ for essentially all samples.

This setting exposes a qualitative difference between the VI field and the MLE gradient. For the clipped-exponential link, $(g^{-1})'(z) = 0$ whenever $z > \log C$. Hence, when the iterates are initialized inside the saturated region, the MLE gradient

$$\nabla \mathcal{L}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^N \left(g^{-1}(\tilde{\boldsymbol{x}}^{i\top}\boldsymbol{\beta}) - y^i\right)\frac{(g^{-1})'(\tilde{\boldsymbol{x}}^{i\top}\boldsymbol{\beta})}{g^{-1}(\tilde{\boldsymbol{x}}^{i\top}\boldsymbol{\beta})}\tilde{\boldsymbol{x}}^i$$

Table 3: Mean squared error of the VI estimator and MLE across link functions and iterations $k$ with dense parameter $\boldsymbol{\beta}^\star = d^{-1/2}(1, \ldots, 1)$. For each $(\text{Link}, k, d, N)$ combination, the smaller error between the two estimators is highlighted in bold. The values in parentheses are standard deviations across 1000 independent repetitions.

| Link, $k$ | $d$ | $N = 100$ | | $N = 200$ | | $N = 500$ | | $N = 1000$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | VI | MLE | VI | MLE | VI | MLE | VI | MLE |
| log $k = 20$ | 10 | .241 (.105) | .241 (.105) | .108 (.057) | .108 (.057) | .025 (.014) | .025 (.014) | .007 (.004) | .007 (.004) |
| | 20 | .384 (.106) | .384 (.106) | .232 (.074) | .232 (.074) | .079 (.030) | .079 (.030) | .023 (.009) | .023 (.009) |
| | 50 | .589 (.082) | .589 (.082) | .432 (.071) | .432 (.071) | .226 (.046) | .226 (.046) | .100 (.025) | .100 (.025) |
| | 100 | .719 (.061) | .719 (.061) | .586 (.062) | .586 (.062) | .379 (.049) | .379 (.049) | .221 (.032) | .221 (.032) |
| log $k = 50$ | 10 | .128 (.069) | .128 (.069) | .048 (.027) | .048 (.027) | .013 (.007) | .013 (.007) | .006 (.003) | .006 (.003) |
| | 20 | .258 (.088) | .258 (.088) | .124 (.049) | .124 (.049) | .033 (.012) | .033 (.012) | .013 (.004) | .013 (.004) |
| | 50 | .485 (.080) | .485 (.080) | .307 (.063) | .307 (.063) | .120 (.030) | .120 (.030) | .045 (.011) | .045 (.011) |
| | 100 | .655 (.065) | .655 (.065) | .486 (.060) | .486 (.060) | .249 (.039) | .249 (.039) | .118 (.021) | .118 (.021) |
| log $k = 100$ | 10 | .104 (.059) | .104 (.059) | .039 (.021) | .039 (.021) | .012 (.006) | .012 (.006) | .006 (.003) | .006 (.003) |
| | 20 | .221 (.078) | .221 (.078) | .101 (.039) | .101 (.039) | .028 (.010) | .028 (.010) | .012 (.004) | .012 (.004) |
| | 50 | .452 (.084) | .452 (.084) | .269 (.059) | .269 (.059) | .096 (.025) | .096 (.025) | .037 (.008) | .037 (.008) |
| | 100 | .633 (.064) | .633 (.064) | .448 (.059) | .448 (.059) | .214 (.035) | .214 (.035) | .095 (.017) | .095 (.017) |
| log $k = 200$ | 10 | .096 (.053) | .096 (.053) | .037 (.020) | .037 (.020) | .012 (.006) | .012 (.006) | .006 (.003) | .006 (.003) |
| | 20 | .207 (.075) | .207 (.075) | .093 (.037) | .093 (.037) | .027 (.010) | .027 (.010) | .013 (.004) | .013 (.004) |
| | 50 | .442 (.081) | .442 (.081) | .259 (.057) | .259 (.057) | .089 (.022) | .089 (.022) | .036 (.008) | .036 (.008) |
| | 100 | .626 (.069) | .626 (.069) | .439 (.059) | .439 (.059) | .207 (.035) | .207 (.035) | .090 (.016) | .090 (.016) |
| clipped exp. $k = 20$ | 10 | **.593** (.068) | .595 (.066) | **.484** (.054) | .488 (.051) | **.342** (.042) | .346 (.036) | **.241** (.030) | .245 (.024) |
| | 20 | **.696** (.052) | .697 (.051) | **.593** (.047) | .595 (.045) | **.449** (.037) | .452 (.034) | **.340** (.028) | .344 (.024) |
| | 50 | **.800** (.040) | .801 (.039) | **.722** (.035) | .722 (.034) | **.593** (.030) | .595 (.029) | **.484** (.024) | .486 (.023) |
| | 100 | **.865** (.027) | .865 (.027) | **.802** (.026) | .803 (.025) | **.693** (.024) | .694 (.023) | **.594** (.021) | .596 (.020) |
| clipped exp. $k = 50$ | 10 | **.444** (.084) | .456 (.076) | **.334** (.066) | .344 (.056) | **.201** (.048) | .208 (.036) | **.125** (.033) | .128 (.022) |
| | 20 | **.571** (.074) | .580 (.069) | **.447** (.061) | .457 (.055) | **.297** (.044) | .306 (.037) | **.199** (.032) | .206 (.025) |
| | 50 | **.724** (.055) | .729 (.051) | **.608** (.046) | .616 (.042) | **.448** (.038) | .458 (.034) | **.332** (.031) | .342 (.026) |
| | 100 | **.814** (.039) | .818 (.036) | **.722** (.038) | .727 (.035) | **.569** (.033) | .578 (.030) | **.450** (.026) | .459 (.024) |
| clipped exp. $k = 100$ | 10 | **.388** (.090) | .403 (.081) | **.273** (.068) | .285 (.057) | **.156** (.048) | .161 (.034) | **.089** (.033) | .090 (.021) |
| | 20 | **.519** (.080) | .532 (.072) | **.394** (.065) | .409 (.057) | **.239** (.044) | .251 (.037) | **.154** (.033) | .160 (.024) |
| | 50 | **.691** (.060) | .698 (.055) | **.566** (.057) | .578 (.052) | **.392** (.041) | .406 (.036) | **.274** (.032) | .287 (.026) |
| | 100 | **.803** (.047) | .804 (.043) | **.691** (.043) | .698 (.040) | **.519** (.034) | .532 (.031) | **.390** (.029) | .404 (.025) |
| clipped exp. $k = 200$ | 10 | **.377** (.094) | .393 (.083) | **.255** (.070) | .270 (.058) | **.141** (.046) | .145 (.033) | **.080** (.032) | .080 (.020) |
| | 20 | **.502** (.084) | .516 (.075) | **.375** (.068) | .391 (.058) | **.228** (.047) | .239 (.038) | **.142** (.034) | .147 (.024) |
| | 50 | **.685** (.066) | .692 (.060) | **.547** (.055) | .560 (.050) | **.375** (.042) | .391 (.037) | **.259** (.033) | .272 (.027) |
| | 100 | .800 (.048) | **.799** (.043) | **.680** (.046) | .688 (.041) | **.507** (.038) | .521 (.034) | **.376** (.030) | .391 (.026) |
| GMM CDF $k = 20$ | 10 | **.167** (.061) | .459 (.094) | **.076** (.030) | .271 (.069) | **.023** (.010) | .074 (.026) | **.010** (.004) | .017 (.007) |
| | 20 | **.302** (.070) | .610 (.070) | **.166** (.043) | .456 (.068) | **.057** (.017) | .209 (.043) | **.022** (.007) | .073 (.017) |
| | 50 | **.513** (.067) | .754 (.045) | **.348** (.050) | .648 (.044) | **.164** (.026) | .452 (.043) | **.076** (.013) | .267 (.033) |
| | 100 | **.663** (.053) | .827 (.031) | **.511** (.048) | .753 (.032) | **.299** (.031) | .609 (.032) | **.165** (.019) | .451 (.029) |
| GMM CDF $k = 50$ | 10 | **.103** (.045) | .200 (.079) | **.049** (.023) | .079 (.036) | .019 (.009) | **.017** (.008) | .010 (.004) | **.007** (.003) |
| | 20 | **.210** (.066) | .371 (.086) | **.105** (.035) | .201 (.053) | **.039** (.013) | .055 (.018) | .019 (.006) | **.017** (.005) |
| | 50 | **.435** (.074) | .588 (.068) | **.260** (.051) | .426 (.059) | **.104** (.022) | .199 (.035) | **.049** (.010) | .079 (.015) |
| | 100 | **.624** (.063) | .715 (.048) | **.435** (.055) | .586 (.050) | **.213** (.028) | .370 (.038) | **.105** (.015) | .199 (.025) |
| GMM CDF $k = 100$ | 10 | **.098** (.046) | .145 (.066) | **.048** (.023) | .052 (.025) | .019 (.009) | **.014** (.006) | .010 (.005) | **.007** (.003) |
| | 20 | **.200** (.065) | .297 (.082) | **.100** (.033) | .143 (.045) | .039 (.013) | **.037** (.013) | .019 (.006) | **.014** (.005) |
| | 50 | **.424** (.079) | .525 (.072) | **.240** (.049) | .344 (.058) | **.099** (.021) | .141 (.028) | **.048** (.010) | .051 (.011) |
| | 100 | **.625** (.070) | .671 (.056) | **.419** (.056) | .520 (.051) | **.197** (.027) | .292 (.036) | **.098** (.015) | .140 (.020) |
| GMM CDF $k = 200$ | 10 | **.099** (.050) | .131 (.067) | .049 (.023) | **.047** (.023) | .020 (.010) | **.014** (.007) | .010 (.004) | **.007** (.003) |
| | 20 | **.196** (.066) | .273 (.079) | **.097** (.032) | .127 (.042) | .038 (.013) | **.034** (.012) | .019 (.006) | **.014** (.004) |
| | 50 | **.421** (.083) | .507 (.077) | **.241** (.049) | .328 (.057) | **.098** (.020) | .125 (.027) | .049 (.010) | **.047** (.010) |
| | 100 | **.629** (.075) | .660 (.057) | **.421** (.056) | .503 (.052) | **.195** (.029) | .271 (.035) | **.097** (.015) | .127 (.020) |

vanishes. In contrast, the VI field

$$V_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \big(g^{-1}(\tilde{\boldsymbol{x}}^{i\top}\boldsymbol{\beta}) - y^i\big)\tilde{\boldsymbol{x}}^i$$

does not contain this derivative factor and can remain nonzero even in the saturated region.

Figure 4 illustrates this behavior. The top-left panel confirms that the initial predictors lie in the clipped region. The top-right panel shows that MLE remains essentially at its initial point, while VI continues to reduce the squared parameter error. The bottom-left panel shows that the MLE gradient norm is zero throughout the trajectory, whereas the VI field norm remains nonzero. The bottom-right panel confirms the same qualitative behavior averaged over 200 independent replications.

Table 4: Mean squared error of the VI estimator and MLE across link functions and iterations $k$ with sparse parameter $\boldsymbol{\beta}^\star = (2/\sqrt{5}, 1/\sqrt{5}, 0, \ldots, 0)$.

| Link, $k$ | $d$ | $N = 100$ | | $N = 200$ | | $N = 500$ | | $N = 1000$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | VI | MLE | VI | MLE | VI | MLE | VI | MLE |
| softplus $k = 20$ | 10 | **.632** (.078) | .717 (.061) | **.513** (.069) | .619 (.055) | **.339** (.054) | .464 (.046) | **.216** (.037) | .342 (.034) |
| | 20 | **.731** (.064) | .795 (.049) | **.625** (.058) | .711 (.046) | **.463** (.047) | .577 (.038) | **.334** (.036) | .461 (.030) |
| | 50 | **.832** (.043) | .872 (.032) | **.757** (.042) | .815 (.032) | **.631** (.037) | .716 (.028) | **.508** (.030) | .615 (.024) |
| | 100 | **.882** (.031) | .909 (.023) | **.823** (.029) | .865 (.022) | **.726** (.028) | .791 (.021) | **.626** (.027) | .712 (.021) |
| softplus $k = 50$ | 10 | **.469** (.106) | .570 (.088) | **.341** (.080) | .455 (.069) | **.159** (.048) | .268 (.048) | **.078** (.028) | .161 (.032) |
| | 20 | **.585** (.082) | .668 (.066) | **.459** (.078) | .562 (.066) | **.276** (.053) | .393 (.048) | **.164** (.033) | .273 (.033) |
| | 50 | **.748** (.058) | .798 (.046) | **.637** (.062) | .710 (.050) | **.460** (.052) | .563 (.043) | **.324** (.034) | .439 (.030) |
| | 100 | **.827** (.046) | .858 (.036) | **.741** (.047) | .792 (.037) | **.596** (.039) | .677 (.032) | **.465** (.032) | .567 (.027) |
| softplus $k = 100$ | 10 | **.400** (.117) | .498 (.102) | **.258** (.080) | .365 (.075) | **.115** (.042) | .201 (.046) | **.048** (.021) | .105 (.027) |
| | 20 | **.535** (.103) | .617 (.085) | **.401** (.083) | .501 (.070) | **.215** (.051) | .321 (.049) | **.116** (.035) | .204 (.038) |
| | 50 | **.717** (.064) | .766 (.050) | **.584** (.063) | .659 (.052) | **.397** (.048) | .498 (.041) | **.258** (.037) | .365 (.034) |
| | 100 | **.818** (.053) | .843 (.042) | **.706** (.053) | .756 (.043) | **.539** (.042) | .621 (.035) | **.394** (.031) | .495 (.027) |
| softplus $k = 200$ | 10 | **.378** (.111) | .476 (.099) | **.232** (.079) | .335 (.077) | **.100** (.037) | .179 (.043) | **.045** (.019) | .093 (.025) |
| | 20 | **.527** (.106) | .607 (.089) | **.382** (.080) | .479 (.071) | **.200** (.051) | .300 (.050) | **.102** (.029) | .182 (.032) |
| | 50 | **.707** (.075) | .754 (.061) | **.574** (.069) | .647 (.057) | **.375** (.049) | .475 (.043) | **.241** (.034) | .344 (.032) |
| | 100 | **.817** (.057) | .839 (.045) | **.701** (.058) | .750 (.046) | **.521** (.045) | .603 (.038) | **.376** (.036) | .475 (.032) |
| clipped exp. $k = 20$ | 10 | **.596** (.064) | .598 (.062) | **.487** (.057) | .491 (.054) | **.339** (.039) | .343 (.035) | **.241** (.030) | .244 (.024) |
| | 20 | **.694** (.053) | .695 (.051) | **.595** (.042) | .597 (.040) | **.444** (.038) | .448 (.035) | **.341** (.027) | .345 (.024) |
| | 50 | **.800** (.037) | .801 (.036) | **.721** (.036) | .722 (.035) | **.590** (.028) | .592 (.027) | **.485** (.025) | .488 (.023) |
| | 100 | **.867** (.024) | .868 (.024) | **.799** (.025) | .799 (.025) | **.692** (.025) | .693 (.024) | **.595** (.019) | .597 (.019) |
| clipped exp. $k = 50$ | 10 | **.447** (.085) | .457 (.077) | **.325** (.061) | .337 (.051) | **.203** (.045) | .210 (.034) | **.126** (.036) | .130 (.023) |
| | 20 | **.572** (.067) | .581 (.061) | **.445** (.063) | .459 (.056) | **.298** (.045) | .308 (.037) | **.199** (.031) | .206 (.024) |
| | 50 | **.720** (.058) | .726 (.054) | **.611** (.048) | .619 (.045) | **.445** (.041) | .455 (.036) | **.332** (.027) | .342 (.023) |
| | 100 | **.813** (.040) | .816 (.037) | **.724** (.041) | .729 (.038) | **.566** (.032) | .575 (.029) | **.446** (.027) | .457 (.024) |
| clipped exp. $k = 100$ | 10 | **.390** (.093) | .406 (.083) | **.275** (.068) | .287 (.055) | **.156** (.045) | .164 (.033) | .091 (.036) | **.091** (.025) |
| | 20 | **.520** (.082) | .533 (.074) | **.391** (.062) | .407 (.056) | **.241** (.049) | .251 (.038) | **.154** (.034) | .159 (.025) |
| | 50 | **.684** (.054) | .692 (.051) | **.567** (.052) | .578 (.048) | **.389** (.043) | .404 (.037) | **.278** (.032) | .290 (.027) |
| | 100 | **.799** (.046) | .801 (.041) | **.688** (.045) | .696 (.041) | **.521** (.036) | .533 (.032) | **.393** (.030) | .407 (.026) |
| clipped exp. $k = 200$ | 10 | **.384** (.085) | .396 (.076) | **.259** (.072) | .273 (.061) | **.142** (.046) | .148 (.035) | **.078** (.030) | .080 (.019) |
| | 20 | **.506** (.086) | .520 (.078) | **.368** (.065) | .386 (.057) | **.230** (.048) | .239 (.038) | **.142** (.034) | .149 (.024) |
| | 50 | **.676** (.063) | .683 (.059) | **.550** (.054) | .562 (.049) | **.373** (.040) | .389 (.034) | **.265** (.032) | .276 (.026) |
| | 100 | **.798** (.054) | .798 (.048) | **.682** (.047) | .690 (.043) | **.504** (.037) | .518 (.033) | **.374** (.031) | .390 (.027) |
| GMM CDF $k = 20$ | 10 | **.171** (.066) | .467 (.097) | **.075** (.027) | .273 (.068) | **.022** (.010) | .073 (.023) | **.009** (.004) | .017 (.007) |
| | 20 | **.304** (.072) | .608 (.070) | **.164** (.041) | .456 (.065) | **.057** (.017) | .208 (.038) | **.022** (.007) | .074 (.016) |
| | 50 | **.512** (.066) | .756 (.043) | **.348** (.052) | .647 (.048) | **.166** (.026) | .455 (.040) | **.075** (.014) | .266 (.031) |
| | 100 | **.666** (.051) | .828 (.030) | **.515** (.047) | .755 (.034) | **.302** (.031) | .611 (.031) | **.165** (.019) | .453 (.031) |
| GMM CDF $k = 50$ | 10 | **.107** (.054) | .201 (.086) | **.050** (.023) | .084 (.037) | .019 (.009) | **.017** (.008) | .009 (.004) | **.007** (.003) |
| | 20 | **.221** (.072) | .384 (.093) | **.105** (.035) | .199 (.053) | **.039** (.013) | .054 (.018) | .019 (.006) | **.017** (.005) |
| | 50 | **.421** (.079) | .577 (.071) | **.259** (.050) | .429 (.058) | **.104** (.019) | .199 (.033) | **.048** (.009) | .076 (.014) |
| | 100 | **.627** (.070) | .719 (.053) | **.433** (.057) | .585 (.050) | **.212** (.027) | .370 (.035) | **.104** (.015) | .199 (.023) |
| GMM CDF $k = 100$ | 10 | **.095** (.046) | .134 (.063) | **.051** (.026) | .055 (.027) | .019 (.008) | **.014** (.006) | .010 (.005) | **.006** (.003) |
| | 20 | **.201** (.063) | .301 (.087) | **.098** (.033) | .143 (.050) | .038 (.012) | **.037** (.012) | .019 (.006) | **.014** (.004) |
| | 50 | **.421** (.081) | .522 (.071) | **.244** (.049) | .347 (.063) | **.098** (.020) | .141 (.029) | **.049** (.011) | .052 (.012) |
| | 100 | **.623** (.068) | .670 (.054) | **.425** (.052) | .522 (.049) | **.199** (.032) | .288 (.037) | **.100** (.015) | .142 (.021) |
| GMM CDF $k = 200$ | 10 | **.096** (.047) | .128 (.064) | .051 (.024) | **.048** (.023) | .021 (.010) | **.014** (.007) | .009 (.004) | **.007** (.003) |
| | 20 | **.192** (.060) | .277 (.082) | **.095** (.033) | .130 (.041) | .039 (.013) | **.033** (.011) | .020 (.006) | **.014** (.005) |
| | 50 | **.415** (.077) | .499 (.071) | **.245** (.050) | .331 (.063) | **.099** (.020) | .131 (.026) | .048 (.010) | **.046** (.010) |
| | 100 | **.629** (.076) | .659 (.059) | **.426** (.061) | .506 (.056) | **.196** (.031) | .273 (.037) | **.098** (.015) | .125 (.019) |

This experiment demonstrates that the plateau pathology is not merely a one-dimensional illustration. In a multidimensional Poisson GLM with a clipped-exponential link, MLE can become completely trapped under a saturated initialization, while VI can still move.

### G.3 Synthetic GAM reconstruction

We construct the GAM reconstruction experiments considered in Section 6. We consider the Poisson and Bernoulli response models. Each additive component $f_j$ is represented using a truncated Legendre basis of degree $K$, with the coefficient of $P_0 \equiv 1$ fixed to zero. This gives the reference-interval centering convention $\int_{-1}^{1} f_j(u)\, du = 0$, rather than the pointwise constraint $f_j(0) = 0$. We use $d = 4$ additive components and $K = 3$ basis functions per component, with covariates independently

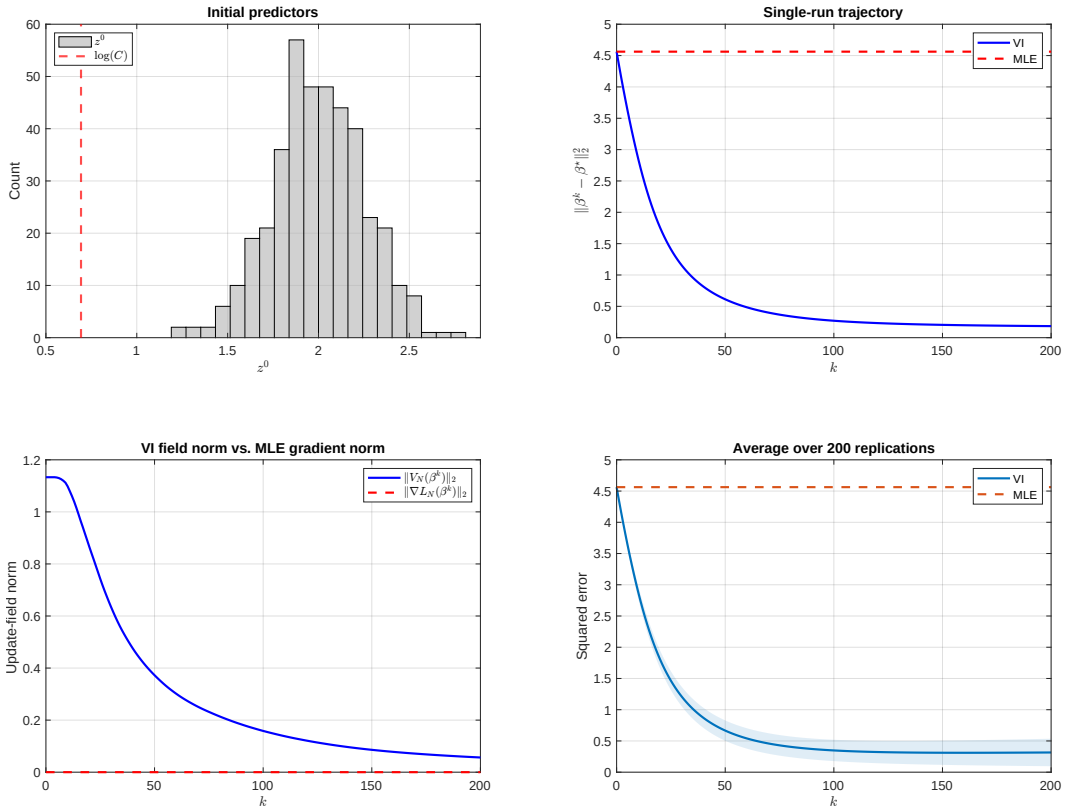



Figure 4: Saturated-initialization stress test for the clipped-exponential Poisson model. Both VI and MLE are initialized in the clipped region. Since the clipped-exponential derivative vanishes in this region, the MLE gradient is zero and the MLE trajectory remains trapped near its initial point. In contrast, the VI field remains nonzero and the VI iterate continues to reduce the squared parameter error. The averaged trajectory over 200 replications shows the same qualitative behavior.

sampled from $U[-1, 1]$. Each $\beta_{jk}$ is drawn from a standard normal distribution, and the coefficient vector is then scaled so that $\|\boldsymbol{\beta}\|_2 = 1.5$.

For Poisson models, we consider the canonical log link as well as softplus, clipped exponential, and Gaussian-mixture CDF links. For Bernoulli models, we consider the canonical logistic link together with the non-canonical arctangent link, whose inverse is given by

$$g^{-1}(z) = \frac{1}{2} + \frac{1}{\pi}\arctan(z).$$

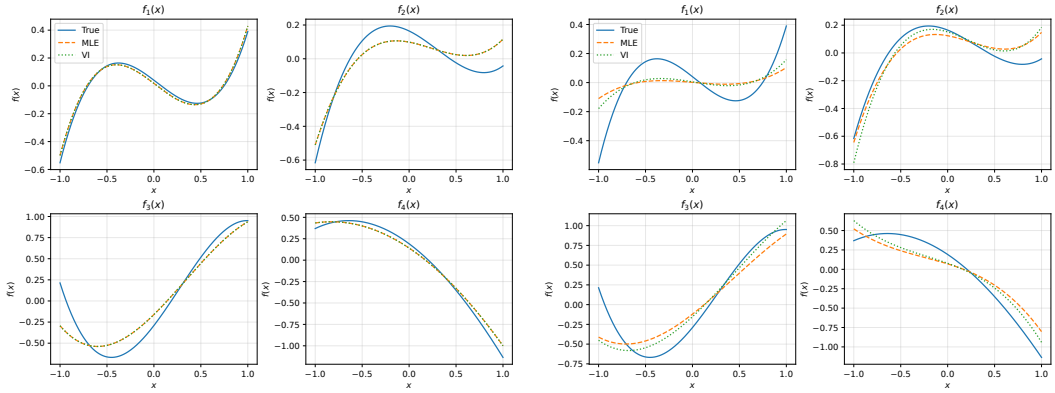
Figures 5 and 6 compare the recovered additive components with the ground truths, while Figures 7 and 8 report the squared error of the parameter estimates and the average $L^2$ errors of the additive components over iterations for the Poisson and Bernoulli models, respectively. For non-canonical links in both Poisson and Bernoulli GAMs, the VI estimator often yields slightly smoother optimization trajectories and modestly improved reconstruction of the additive components compared to MLE, although the final errors are typically of similar magnitude. These experiments illustrate that the proposed VI framework extends naturally to GAMs through basis representations and remains numerically stable in additive models beyond linear predictors.

### G.4 Real-data bike-sharing benchmark

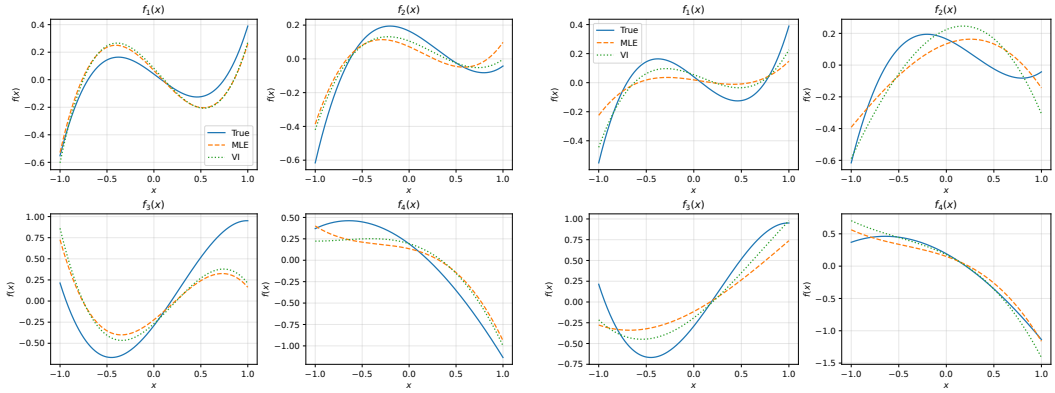
We provide additional details for the real-data experiment in Section 7.3. We compare the log, softplus, and clipped exponential links

$$g^{-1}(z) = \exp(z), \quad g^{-1}(z) = \log(1 + \exp(z)), \quad g^{-1}(z) = \min\{\exp(z), C\},$$

where the cap $C$ for the clipped exponential link is selected by validation Poisson deviance.

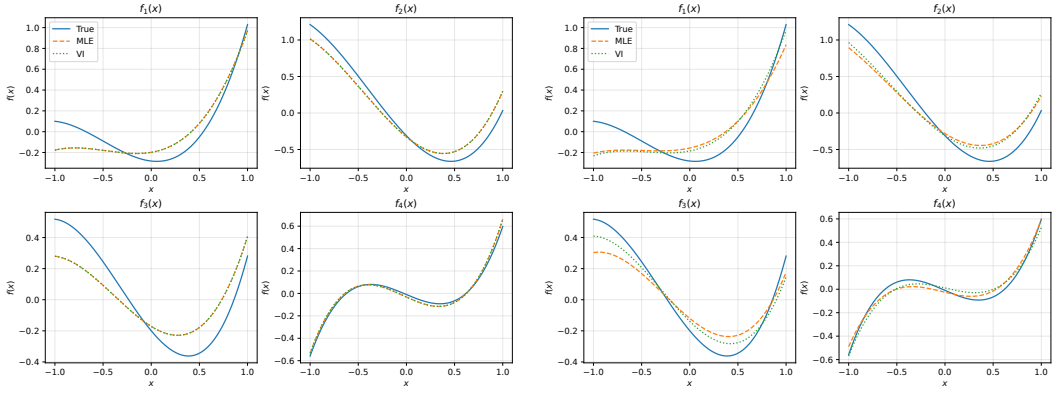



(a) log (b) softplus




(c) clipped exp. (d) GMM CDF

Figure 5: Reconstruction of additive components in GAMs under Poisson distribution.

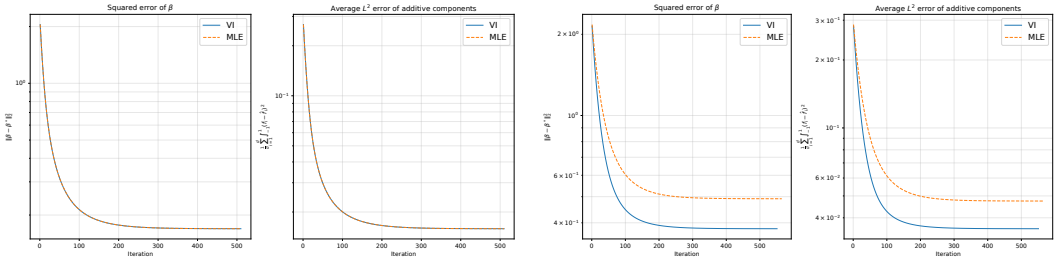



(a) logistic (b) arctangent

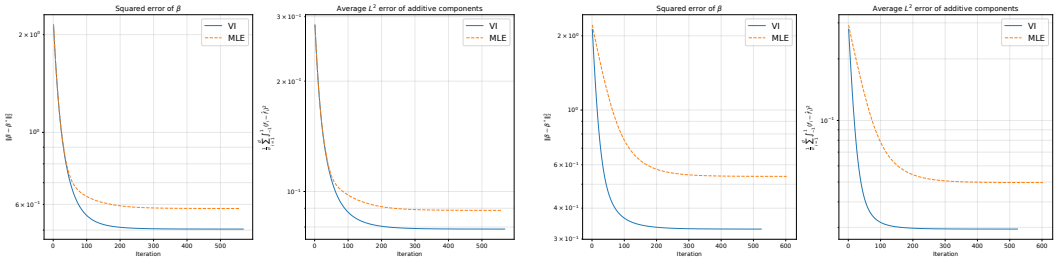
Figure 6: Reconstruction of additive components in GAMs under Bernoulli distribution.

The chronological train/validation/test split contains 10,427/3,475/3,477 observations. For the fixed-budget softplus fits, the step size and checkpoint are selected by validation deviance from 250-step trajectories. Mean Poisson deviance uses the predicted-mean floor $\max\{\mu_i, 10^{-12}\}$. These data and metric conventions also apply to the converged comparisons in Appendix G.5.

Figure 9 shows runtime versus validation Poisson deviance for the GLM and GAM designs. Under the log link, VI and MLE have nearly identical trajectories. Under the softplus link, VI reaches a

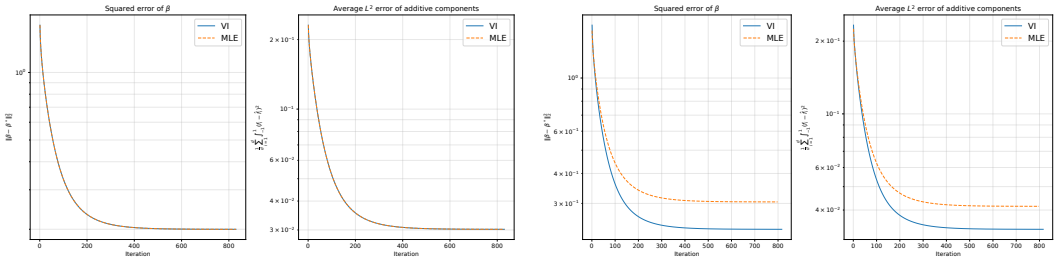



(a) log (b) softplus



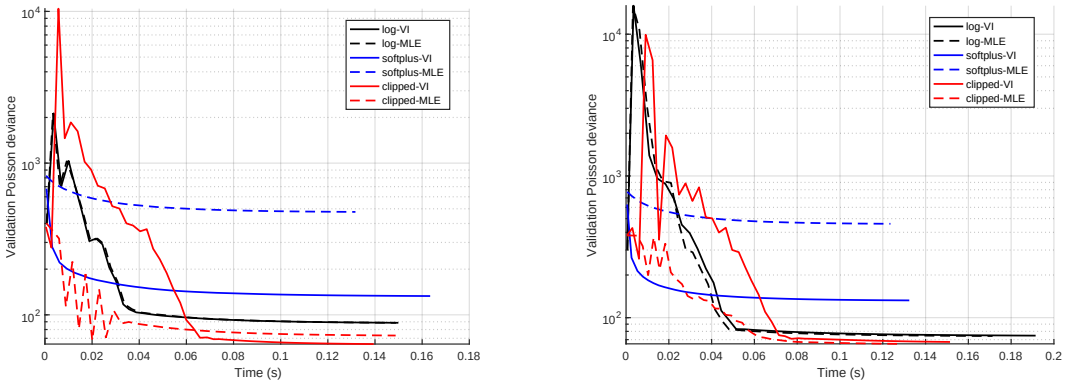

(c) clipped exp. (d) GMM CDF

Figure 7: Squared error of parameter estimates and the average $L^2$ error of the additive components (Poisson response).




(a) logistic (b) arctangent

Figure 8: Squared error of parameter estimates and the average $L^2$ error of the additive components (Bernoulli response).




(a) GLM (b) GAM, $K = 3$

Figure 9: Runtime versus validation Poisson deviance.

much lower validation deviance than MLE at comparable runtime. The clipped exponential link is more mixed.

Table 5: Test metrics on the bike-sharing dataset.

| Model | Link | Test Dev. VI | Test Dev. MLE | RMSE VI | RMSE MLE | MAE VI | MAE MLE |
|---|---|---|---|---|---|---|---|
| GLM | log | 59.43 | 59.43 | 125.30 | 125.30 | 92.45 | 92.45 |
| GLM | softplus | **133.99** | 453.85 | **187.78** | 280.89 | **135.61** | 201.89 |
| GLM | clipped-exp | **51.90** | 59.84 | **128.16** | 128.17 | **88.65** | 94.05 |
| GAM, $K = 3$ | log | 55.57 | 56.61 | 125.65 | 126.88 | 88.97 | 89.82 |
| GAM, $K = 3$ | softplus | **129.48** | 428.88 | **185.03** | 277.65 | **132.33** | 199.37 |
| GAM, $K = 3$ | clipped-exp | 55.32 | **50.15** | 133.50 | **125.49** | 93.37 | **86.64** |

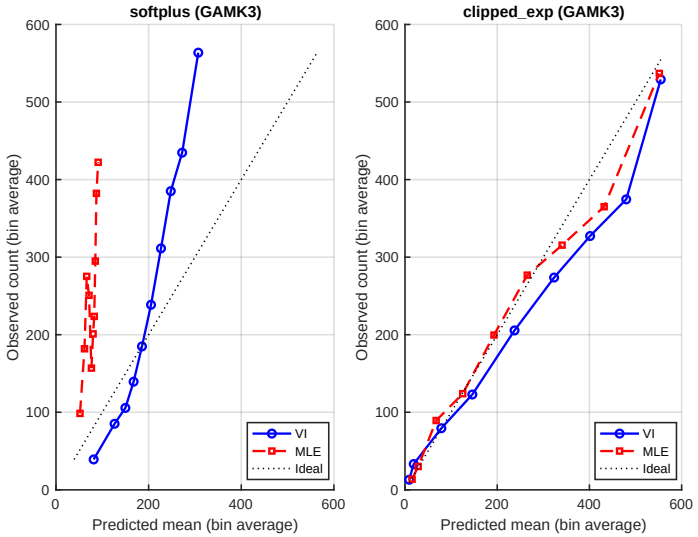



Figure 10: Test calibration plots (GAM, $K = 3$). The diagonal line denotes perfect calibration.

Table 5 reports the test Poisson deviance, RMSE, and MAE. The softplus link shows the clearest advantage for VI in both GLM and GAM designs. For the clipped exponential link, VI is better in the GLM design, while MLE is slightly better in the GAM design.

Figure 10 reports test-set calibration for the GAM design under the softplus and clipped exponential links. For each bin, the $x$-axis shows the average predicted mean and the $y$-axis shows the average observed count. The softplus case shows a more stable calibration pattern for VI, whereas the clipped exponential case is more comparable between the two methods.

### G.5 Numerically converged softplus comparisons

We compare fixed-budget and numerically converged estimators under the softplus link, for both simulation setting of Section 7.1 and bike-sharing experiment of Section 7.3. The simulation uses the dense, no-intercept model of Section 7.1 with $d = 20$, $N = 400$, and 1,000 independent paired repetitions. Within each repetition, all four methods are fitted to the same generated sample. The bike-sharing data and evaluation metrics follow Appendix G.4.

For the numerically converged fits, both unpenalized fits start at zero and are accepted only when

$$\frac{\|F_N(\hat{\boldsymbol{\beta}})\|_\infty}{1 + \|\hat{\boldsymbol{\beta}}\|_\infty} \le 10^{-8}, \qquad F_N = V_N \text{ (VI)}, \quad F_N = \nabla \mathcal{L}_N \text{ (MLE)},$$

where both fields are sample averages. Synthetic VI uses SciPy's hybrid root solver with an analytic Jacobian. Bike VI uses damped Newton updates with an analytic Jacobian and a squared-residual line search. MLE uses unconstrained SciPy L-BFGS-B with an analytic gradient.

Table 6 shows that the large fixed-budget MLE disadvantage disappears at convergence in the correctly specified simulation, where MLE attains slightly lower parameter MSE than VI. This is consistent

with the asymptotic efficiency comparison under correct specification in Appendix E. On bike sharing, converged MLE substantially narrows the gap, but converged VI retains lower test deviance despite MLE's lower training NLL.

Table 6: Fixed-budget and numerically converged softplus comparisons. For simulation, entries are mean squared parameter errors (SDs). Training NLL for bike sharing is the mean of $\mu_i - y_i \log \mu_i$, omitting $\log(y_i!)$. Fixed budgets are $k = 200$ for simulation and $k = 250$ for bike sharing.

| | Fixed iteration budget | | Numerically converged | |
|---|---|---|---|---|
| Setting / metric | VI | MLE-GD | VI root | MLE (L-BFGS-B) |
| Simulation: parameter MSE | 0.241 (0.054) | 0.343 (0.051) | 0.175 (0.056) | **0.170** (0.056) |
| Bike GLM: test deviance | 133.99 | 453.85 | **76.63** | 100.46 |
| Bike GAM: test deviance | 129.48 | 428.88 | **74.67** | 96.52 |
| Bike GLM: training NLL | $-603.88$ | $-541.76$ | $-528.13$ | $-617.20$ |
| Bike GAM: training NLL | $-600.48$ | $-547.67$ | $-524.09$ | $-617.90$ |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main claims, which are supported by both theories and extensive experiments conducted.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a dedicated "Discussion" section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper states assumptions for the main theoretical results and provides full proofs in appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The algorithm, datasets, metrics and baselines details are described in the paper and appendix, which is enough to understand how the main results were produced.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The experiment code, environment requirements, execution instructions, and Bike Sharing data are publicly available at `https://github.com/jonghyeoki/vi-glm-estimation`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting is described in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All the tables report mean with standard deviation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper is primarily theoretical, and reproducing its results does not require substantial computational resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper studies basic statistical estimation problems. Potential negative impacts are limited because our work is foundational and not tied to a deployed application.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We developed our own code and simulated datasets, and we correctly credited and cited the real-world datasets we used and the baselines we compared.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are introduced in this paper. The paper is primarily theoretical, and the numerical experiments are based on data according to the experimental settings described in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.