A Projection to the Feasible Set

As noted in the Section 3, projecting some value of θ to $\mathcal{R}_\alpha$ amounts to ensuring that $\theta^\top x$ is in some interval $[r, s]$ for some $r, s \in \mathbb{R}$. The values of r and s can be analytically computed on the type of noise and the truncation interval $[a, b]$. To show that we can efficiently project to $\mathcal{R}_\alpha$, however, we demonstrate that we can efficiently project a given θ_0 to any set of the form $\{\theta \mid \max_{x\in\mathcal{X}} x^\top\theta \leq C\}$.

Note that this projection is equivalent to solving the following convex program for a given θ_0:

$$\begin{aligned} &\min_\theta \|\theta - \theta_0\|^2 \\ &\theta^\top x < C \qquad\qquad\qquad\qquad \forall \qquad x \in \mathcal{X}. \end{aligned}$$

Note that this program does have infinite constraints, but it is straightforward to craft an efficient *separation oracle* for the set defined by these constraints, simply by maximizing $\theta^\top x$ constrained to $\mathcal{X}$. We can thus exploit the ellipsoid algorithm [Gup15] in order to efficiently solve the projection step.

Handling approximation. The above methodology provides an efficient routine for projecting parameter estimates θ into $\mathcal{R}_\alpha$ provided it is possible to perform exact arithmetic on real numbers, and solve $\max_{x\in\mathcal{X}} \theta^\top x$ exactly. In reality, we solve the above programs approximately using, e.g., the central-cut ellipsoid algorithm (c.f. [GLS80] Section 3.2 and references therein), which only requires a lower bound on the volume of the constraint set (which in fact follows straightforwardly from the bound $B = \max_{x\in\mathcal{X}} \|x\|$ and the Cauchy-Schwarz inequality). In the end, we end up with an approximate projection algorithm which does not discernably change the overall runtime of SGD or SNGD.

B Omitted Proofs for Probit Regression

B.1 Second derivative for Truncated Probit Regression

Here we give a derivation of $\nabla_\theta^2 \ell(\theta; y)$ which is used to show the strong concavity of the log-likelihood.

$$\ell(\theta; y) = y \cdot \log\left(\int_0^b \mathcal{N}\left(z; \theta^\top x\right) dz\right) + (1-y) \cdot \log\left(\int_a^0 \mathcal{N}\left(z; \theta^\top x\right) dz\right) - \log\left(\int_a^b \mathcal{N}\left(z; \theta^\top x\right) dz\right). \tag{16}$$

In the following, we make use of the change of variables $\mu = \theta^\top x$. Then, we make note of the following identity:

$$\frac{d}{d\mu} \log\left[\int_{z\in S} \mathcal{N}(z;\mu)\ dz\right] = \frac{\int_{z\in S} (\mu - z)\cdot\mathcal{N}(z;\mu)\ dz}{\int_{z\in S}\mathcal{N}(z;\mu)\ dz}$$

Applying this yields the gradient:

$$\begin{aligned} \nabla_\theta \ell(\theta; x, y) &= \frac{d\ell(\theta; x, y)}{d\mu} \cdot x \\ &= \left[y \cdot \frac{\int_0^b (\mu - z)\cdot\mathcal{N}(z;\mu)\ dz}{\int_0^b \mathcal{N}(z;\mu)\ dz} + (1-y)\cdot\frac{\int_a^0 (\mu - z)\cdot\mathcal{N}(z;\mu)\ dz}{\int_a^0 \mathcal{N}(z;\mu)\ dz} - \frac{\int_a^b (\mu - z)\cdot\mathcal{N}(z;\mu)\ dz}{\int_a^b \mathcal{N}(z;\mu)\ dz}\right]\cdot x \\ &= \left[-y \cdot \frac{\int_0^b z\cdot\mathcal{N}(z;\mu)\ dz}{\int_0^b \mathcal{N}(z;\mu)\ dz} - (1-y)\cdot\frac{\int_a^0 z\cdot\mathcal{N}(z;\mu)\ dz}{\int_a^0 \mathcal{N}(z;\mu)\ dz} + \frac{\int_a^b z\cdot\mathcal{N}(z;\mu)\ dz}{\int_a^b \mathcal{N}(z;\mu)\ dz}\right]\cdot x \end{aligned}$$

To get the second derivative, we make use of the following:

$$\begin{aligned} \frac{d}{d\mu}\frac{\int_{z\in S} z\cdot\mathcal{N}(z;\mu)\ dz}{\int_{z\in S}\mathcal{N}(z;\mu)\ dz} &= \frac{\left(\int_{z\in S} -z^2\cdot\mathcal{N}(z;\mu)dz\right)}{\left(\int_{z\in S}\mathcal{N}(z;\mu)dz\right)^2} + \frac{\left(\int_{z\in S} z\cdot\mathcal{N}(z;\mu)dz\right)^2}{\left(\int_{z\in S}\mathcal{N}(z;\mu)dz\right)^2} \\ &= -\mathrm{Var}_{z \overset{z\in S}{\sim} \mathcal{N}(\cdot;\mu)}[z] \end{aligned}$$

An application of the above identity yields

$$\nabla^2_\theta \ell(\theta; x, y) = \left[y \cdot \operatorname{Var}_{z \overset{z\in[0,b]}{\sim} \mathcal{N}(\cdot;\mu)}[z] + (1-y)\cdot \operatorname{Var}_{z \overset{z\in[a,0]}{\sim} \mathcal{N}(\cdot;\mu)}[z] - \operatorname{Var}_{z \overset{z\in[a,b]}{\sim} \mathcal{N}(\cdot;\mu)}[z] \right] \cdot xx^\top,$$

which is precisely (11).

B.2 Proofs of Lemmata

Lemma 1. *Let θ_0 be the probability density function of the normal distribution with zero mean and unit variance, truncated to a half-open interval (r, ∞). Now, we define the function $V(s)$ to be the variance of the left-truncated distribution when* also *right-truncated at $s > r$, that is, $V(s) = Var_{x\sim\mathcal{N}(0,1)}[x|r \leq x \leq s]$. Then, for any $b_1, b_2, \gamma \in \mathbb{R}$ such that $s > b_2 > b_1 > r$ and $P_{x\sim\mathcal{N}(0,1)}([b_1, b_2]) > \gamma > 0$, we have that*

$$V(b_2) - V(b_1) > poly(\gamma, b_2 - b_1).$$

Proof. For $i \in \{1, 2, 3\}$, we define θ_i recursively as:

$$\theta_i(x) = \int_{-\infty}^x \theta_{i-1}(z)\ dz = \int_a^x \theta_{i-1}(z)\ dz. \tag{17}$$

For convenience, we write $\theta_i'(x)$ to denote the derivative of θ_i with respect to its argument at x, so that $\theta_i'(x) = \theta_{i-1}(x)$.

Now, we define the function $V(b)$ to be the variance of the left-truncated distribution when *also* right-truncated at $b > a$. Via integration by parts one finds that [Bur96]:

$$V(b_2) - V(b_1) = 2 \cdot \int_{b_1}^{b_2} \frac{\theta_0(z)}{\theta_1(z)^3} \cdot \left(\theta_3'^2(z) - \theta_3''(z)\cdot\theta_3(z)\right)\ dz. \tag{18}$$

Our goal is to show that for $b_2 > b_1$, (18) is lower bounded. First, note that the quantity is clearly non-negative. In particular, by a direct application of the Prekopa-Leindler inequality [Pré73] we find that θ_i log-concave $\implies$ θ_{i+1} log-concave. Since θ_0 is log-concave[2] on $[b_1, b_2]$, so is $\theta_3(\cdot)$, making (18) non-negative.

To show that (18) is lower bounded, we lower bound the integrand on a subset of the interval $[b_1, b_2]$. We show this inductively—for a given i, suppose that $\theta_i(\cdot)$ is c-strongly log-concave ($c > 0$) on an interval $[\alpha_i, \beta_i] \subset [b_1, b_2]$. Concretely, this means that for any $x, t \in [\alpha_i, \beta_i]$ with $x > t$,

$$\frac{\theta_i'(x)}{\theta_i(x)} \leq \frac{\theta_i'(t)}{\theta_i(t)} - c\cdot(x-t) \tag{19}$$

Note that (19) holds without the $c(x-t)$ term on all of $[b_1, b_2]$ due to ordinary log-concavity. To prove (19) we need to start with an $[\alpha_0, \beta_0]$ such that θ_0 is lower bounded by ρ on $[\alpha_0, \beta_0]$, $\beta_0 - \alpha_0 \geq \tau$ and $P_{x\sim\mathcal{N}(0,1)}([\alpha_0, \beta_0]) > \gamma'$.

To find such an interval we start from the property $P_{x\sim\mathcal{N}(0,1)}([b_1, b_2]) > \gamma$. It is clear that the interval with the smallest values of θ_0 and total mass γ is the interval $[r, \infty)$ for some r such that $\int_r^\infty \theta_0(x)dx = \gamma$. If $r \leq 2$, i.e. then we can set $\alpha_0 = 2$, $\beta_0 = 10$ and we have that $\rho = O(1)$, $\tau = O(1)$ and $\gamma' = O(1)$, so we can assume that $r > 2$. In this case we know from standard bounds of the error function that $\exp(-r^2) \leq \int_r^\infty \theta_0(x)dx \leq \exp(-r^2/2)$ and hence we know that $r \leq \sqrt{2\ln(1/\gamma)}$. We can then set $\alpha_0 = \sqrt{2\ln(1/\gamma)}$ and $\beta_0 = 4\sqrt{\ln(1/\gamma)}$ and using the bounds on the error function for $r > 2$ we get $\rho \geq \text{poly}(\gamma)$, $\tau \geq \text{poly}(\gamma)$ and $\gamma' \geq \text{poly}(\gamma)$.

Before we continue we also define B to be an upper bound on θ_2 which we will bound precisely later. Now, for

[2] This is clear since θ_0 is just a scaled version of the Gaussian PDF, which is strongly log-concave

any $x \in [\alpha_i, \beta_i]$:

$$\begin{aligned}
\frac{\theta_i'(x)}{\theta_i(x)}\theta_{i+1}(x) &= \frac{\theta_i'(x)}{\theta_i(x)}\left(\int_{-\infty}^{\alpha}\theta_i(t)dt + \int_{\alpha}^{x}\theta_i(t)dt\right) \qquad (20)\\
&\leq \int_{-\infty}^{\alpha}\frac{\theta_i'(t)}{\theta_i(t)}\theta_i(t)dt + \int_{\alpha}^{x}\left(\frac{\theta_i'(t)}{\theta_i(t)} - c\cdot(t-x)\right)\theta_i(t)\ dt \qquad (21)\\
&\leq \theta_i(x) - c\cdot\int_{\alpha}^{x}(t-x)\cdot\theta_i(t)\ dt \qquad (22)\\
\theta_i'(x)\theta_{i+1}(x) &\leq \theta_i(x)^2 - c\cdot\theta_i(x)\cdot\int_{\alpha}^{x}(t-x)\cdot\theta_i(t)\ dt \qquad (23)\\
\frac{\theta_i'(x)\theta_{i+1}(x)-\theta_i(x)^2}{\theta_{i+1}(x)^2} &\leq -\frac{c\cdot\theta_i(x)}{\theta_{i+1}(x)^2}\int_{\alpha}^{x}(t-x)\cdot\theta_i(t)\ dt \qquad (24)\\
& \qquad (25)
\end{aligned}$$

We begin with $i = 0$ and $c = 1$ (since $\frac{\partial^2}{\partial x^2}\log\phi(x) = 1$). We define $\alpha_i = \alpha_{i-1} + \varepsilon$ and $\beta_i = \beta_{i-1}$, such that for all $x \in [\alpha_1, \beta_1]$:

$$\frac{\theta_1^{''}(x)\theta_1(x) - \theta_1^{'}(x)^2}{\theta_1(x)^2} \leq -\frac{1}{2}\rho^2\cdot\varepsilon^2, \tag{26}$$

and so θ_1 is in fact strongly log-concave on the interval $[\alpha_1, \alpha_1]$. Also note that by definition, $\theta_1 \geq \varepsilon\cdot\rho$ on this interval (and in turn, $\theta_2 \geq \varepsilon^2\cdot\rho$ on $[\alpha_2, \beta_2]$). Iterating again on the interval $[\alpha_2, \beta_1]$ then $[\alpha_3, \beta_3]$ yields:

$$\begin{aligned}
\frac{\theta_2^{''}(x)\theta_2(x) - \theta_2^{'}(x)^2}{\theta_2(x)^2} &\leq -\frac{\rho^2\cdot\varepsilon^2\cdot(\rho\cdot\varepsilon)^2\cdot\varepsilon^2}{4\cdot\theta_2(x)^2} \leq -\frac{\rho^4\cdot\varepsilon^6}{4\cdot B^2} \qquad (27)\\
\theta_3^{''}(x)\theta_3(x) - \theta_3^{'}(x)^2 &\leq \left(\frac{\rho^4\cdot\varepsilon^6}{4\cdot B^2}\right)\frac{(\varepsilon^2\rho)^2\varepsilon^2}{2} = -\frac{\rho^6\cdot\varepsilon^{12}}{8\cdot B^2}. \qquad (28)
\end{aligned}$$

Observe that setting $\varepsilon = \frac{1}{4}(\beta_0 - \alpha_0)$ allows us to lower bound the difference in variance:

$$\begin{aligned}
V(b_2) - V(b_1) &= 2\cdot\int_{b_1}^{b_2}\frac{\theta_0(z)}{\theta_1(z)^3}\cdot\left(\theta_3'^2(z) - \theta_3''(z)\cdot\theta_3(z)\right)\ dz \qquad (29)\\
&\geq 2\cdot\int_{\alpha_0}^{\beta_0}\frac{\theta_0(z)}{\theta_1(z)^3}\cdot\left(\theta_3'^2(z) - \theta_3''(z)\cdot\theta_3(z)\right)\ dz \qquad (30)\\
&\geq 2\cdot\int_{\frac{3}{4}\alpha_0+\frac{1}{4}\beta_0}^{\beta_0}\theta_0(z)\cdot\left(\theta_3'^2(z) - \theta_3''(z)\cdot\theta_3(z)\right)\ dz \qquad (31)\\
&\geq \text{poly}\left(\rho, (b_2 - b_1), \frac{1}{B}\right). \qquad (32)
\end{aligned}$$

Where both inequalities follow from the positiveness of the integrand. To conclude the proof, it suffices to note that by the construction of α_0, β_0 it holds that $\rho = \Theta(\text{poly}(\gamma))$, and $\max(\alpha_0, \beta_0) = \Theta(\text{poly}(\ln(1/\gamma)))$. Finally, we can compute

$$\theta_2(x) = \frac{x\cdot(\Phi(x) - \Phi(a)) + (\phi(x) - \phi(a))}{\int_a^\infty \phi(z-\mu)\ dz} \leq O(x),$$

which means that $B \in \text{poly}(1/\gamma)$, proving the desired statement. □

C Omitted Proofs for Logistic Regression

C.1 Definition of Strong Local Quasi-Convexity

Here we present the definition of Strong Local Quasi-Convex (SLQC) functions, for which Hazan, Levy, and Shalev-Shwartz [HLS15] provides stochastic convergence guarantees.

Definition 3 (SLQC functions)**.** *Let $x, z \in \mathbb{R}^d, \kappa, \varepsilon > 0$. We say that $f : \mathbb{R}^d \to R$ is (ε, κ, z)-Strictly-Locally-Quasi-Convex (SLQC) in x, if at least one of the following applies:*

1. *$f(x) - f(z) \leq \varepsilon$.*
2. *$\|\nabla f(x)\| > 0$, and for every $y \in \mathcal{B}(z, \varepsilon/\kappa)$ it holds that $\langle \nabla f(x), y - x \rangle \leq 0$.*

C.2 Closed-form Gradient for Logistic Regression

We begin with the following form of the gradient of the population log-likelihood:

$$\nabla_\theta \bar{\ell}(\theta; \theta_*) = 2 \cdot \left(\left[\sum_{y \in \{0,1\}} p^*_{(x,y)} \cdot \mathbb{E}_{z \overset{\phi}{\sim} f_\ell(\cdot;\theta^\top x)} \left[\sigma(z - \theta^\top x) \middle| \mathbf{1}_{z \geq 0} = y \right] \right] - \mathbb{E}_{z \overset{\phi}{\sim} f_\ell(\cdot;\theta^\top x)} \left[\sigma(z - \theta^\top x) \right] \right) \boldsymbol{x}$$

To simplify this expression, consider the cumulative distribution function F of the distribution with density $f_\ell(\cdot; \theta^\top x)$ truncated according to any interval $[a, b]$:

$$F(x) = \frac{\sigma(x - \theta^\top x) - \sigma(a - \theta^\top x)}{\sigma(b - \theta^\top x) - \sigma(a - \theta^\top x)}.$$

This allows us to rewrite the expectations above in terms of F:

$$\begin{aligned} \mathbb{E}_{z \overset{[a,b]}{\sim} f_\ell(\cdot;\theta^\top x)} \left[\sigma \left(z - \theta^\top x \right) \right] &= \mathbb{E}_{z \overset{[a,b]}{\sim} f_\ell(\cdot;\theta^\top x)} \left[\left(\sigma \left(b - \theta^\top x \right) - \sigma \left(a - \theta^\top x \right) \right) \cdot F(z) + \sigma \left(a - \theta^\top x \right) \right] \\ &= \left(\sigma \left(b - \theta^\top x \right) - \sigma \left(a - \theta^\top x \right) \right) \cdot \mathbb{E}_{z \overset{[a,b]}{\sim} f_\ell(\cdot;\theta^\top x)} \left[F(z) \right] + \sigma \left(a - \theta^\top x \right) \\ &= \frac{1}{2} \left(\sigma \left(b - \theta^\top x \right) + \sigma \left(a - \theta^\top x \right) \right), \end{aligned}$$

where the last equality is due to the fact that the distribution of z under its own CDF is the uniform distribution (and thus has expectation $\frac{1}{2}$). Now, note that when $\phi(z) = \mathbf{1}_{z \in [a,b]}$ (which is the case we consider here), all of the expectation terms in the population gradient can be expressed in this way. In particular, we have:

$$\nabla = \left(p^*_{(x,1)} \cdot \left[\sigma \left(b - \theta^\top x \right) + \sigma \left(- \theta^\top x \right) \right] + p^*_{(x,0)} \cdot \left[\sigma \left(- \theta^\top x \right) + \sigma \left(a - \theta^\top x \right) \right] - \left[\sigma \left(b - \theta^\top x \right) + \sigma \left(a - \theta^\top x \right) \right] \right) \boldsymbol{x}.$$

We can simplify this by observing that $p^*_{(x,0)} = 1 - p^*_{(x,1)}$, and thus the preceding simplifies to:

$$\nabla_\theta \bar{\ell}(\theta; \theta_*) = \left(\sigma \left(- \theta^\top x \right) - p^*_{(x,1)} \cdot \sigma \left(a - \theta^\top x \right) - (1 - p^*_{(x,1)}) \cdot \sigma \left(b - \theta^\top x \right) \right) \boldsymbol{x}. \tag{33}$$

Finally, we can write $p^*_{(x,1)}$ as:

$$p^*_{(x,1)} = \frac{\sigma(b - \theta_*^\top x) - \sigma(-\theta_*^\top x)}{\sigma(b - \theta_*^\top x) - \sigma(a - \theta_*^\top x)}.$$

Substituting this into (33) yields:

$$\begin{aligned} \nabla_\theta \bar{\ell}(\theta; \theta_*) &= \left(\sigma \left(- \theta^\top x \right) - p^*_{(x,1)} \cdot \sigma \left(a - \theta^\top x \right) - (1 - p^*_{(x,1)}) \cdot \sigma \left(b - \theta^\top x \right) \right) \boldsymbol{x} && (34) \\ &= \left(\sigma \left(- \theta^\top x \right) - \sigma \left(b - \theta^\top x \right) - p^*_{(x,1)} \cdot \left(\sigma \left(a - \theta^\top x \right) - \sigma \left(b - \theta^\top x \right) \right) \right) \boldsymbol{x} && (35) \\ &= \left(\sigma \left(a - \theta^\top x \right) - \sigma \left(b - \theta^\top x \right) \right) \left(\frac{\sigma \left(b - \theta^\top x \right) - \sigma \left(- \theta^\top x \right)}{\sigma \left(b - \theta^\top x \right) - \sigma \left(a - \theta^\top x \right)} - p^*_{(x,1)} \right) \boldsymbol{x} && (36) \\ &= \left(\sigma \left(a - \theta^\top x \right) - \sigma \left(b - \theta^\top x \right) \right) \left(\frac{\sigma \left(b - \theta^\top x \right) - \sigma \left(- \theta^\top x \right)}{\sigma \left(b - \theta^\top x \right) - \sigma \left(a - \theta^\top x \right)} - \frac{\sigma(b - \theta_*^\top x) - \sigma(-\theta_*^\top x)}{\sigma(b - \theta_*^\top x) - \sigma(a - \theta_*^\top x)} \right) \boldsymbol{x}. && (37) \end{aligned}$$

This concludes the derivation.

C.3 Local quasi-convexity of the population likelihood

Lemma 3. *The population log-likelihood from logistic regression truncated to the interval $[a, b]$ is strictly quasi-concave—in particular, we have that for any $\theta \in \mathcal{R}_\alpha$:*

$$\langle \bar{\ell}(\theta; \theta_*), \theta - \theta_* \rangle \leq -\frac{\alpha^2 \cdot \varepsilon^2}{4 \cdot B^2} \cdot \frac{(1-e^a)(e^b-1)}{e^b - e^a},$$

where B is an upper bound on $\|x\|_2$.

Proof. We first define a function f, whose domain is the set $\mathcal{S} = \{\theta^\top x | \theta \in \mathcal{R}_\alpha, x \in \mathcal{X}\}$:

$$f(x) = \frac{\sigma(b-x) - \sigma(-x)}{\sigma(b-x) - \sigma(a-x)},$$

so that the derivative

$$f'(x) = \frac{(1-e^a)(e^b-1)}{e^b - e^a} \cdot \sigma(x) \cdot (1 - \sigma(x))$$

can be bounded as follows for any $x \in S$:

$$0 \leq \frac{\alpha}{2} \cdot \frac{(1-e^a)(e^b-1)}{e^b - e^a} \leq f'(x) \leq \frac{1}{4} \frac{(1-e^a)(e^b-1)}{e^b - e^a}.$$

Here, the first inequality comes from $a < 0$ and $b > 0$ (otherwise the truncation set removes all the elements from one class and inference is impossible). The second inequality comes from the fact that by construction of $\mathcal{R}_\alpha$, we have that for any $\theta \in \mathcal{R}_\alpha$ and any $x \in \mathcal{X}$,

$$\alpha \leq \min_{y \in \{0,1\}} \mathbb{P}_\theta(z \cdot y \geq 0) \leq \min\left\{\sigma(\theta^\top x), 1 - \sigma(\theta^\top x)\right\}.$$

(Note that the second inequality in the above is due to the fact that we omit terms due to truncation, which only make the bounds stronger). For convenience, we denote the lower and upper bounds on f' as C_0 and C_1 respectively. Now, we can write the gradient of the population log-likelihood in the following more convenient form:

$$\nabla_\theta \bar{\ell}(\theta; \theta_*) = \left(\sigma\left(a - \theta^\top x\right) - \sigma\left(b - \theta^\top x\right)\right)\left(f(\theta^\top x) - f(\theta_*^\top x)\right) \boldsymbol{x} \tag{38}$$

$$\langle \nabla_\theta \bar{\ell}(\theta; \theta_*), \theta - \theta_* \rangle = \left(\sigma\left(a - \theta^\top x\right) - \sigma\left(b - \theta^\top x\right)\right)\left(f(\theta^\top x) - f(\theta_*^\top x)\right)(\theta^\top x - \theta_*^\top x). \tag{39}$$

The intermediate value theorem gives us that $\frac{f(\alpha) - f(\beta)}{\alpha - \beta} > C_0$ for any α, β, and thus we can write (39) as:

$$\begin{aligned}\langle \nabla_\theta \bar{\ell}(\theta; \theta_*), \theta - \theta_* \rangle &= \left(\sigma\left(a - \theta^\top x\right) - \sigma\left(b - \theta^\top x\right)\right)\left(f(\theta^\top x) - f(\theta_*^\top x)\right)(\theta^\top x - \theta_*^\top x) \\ &< C_0 \cdot \left(\sigma\left(a - \theta^\top x\right) - \sigma\left(b - \theta^\top x\right)\right) \cdot (\theta^\top x - \theta_*^\top x)^2\end{aligned}$$

Now, note that the likelihood $\bar{\ell}(\theta; \theta_*)$ is at most $2B$-Lipschitz (where B is an upper bound on $\|x\|$, as the gradient (c.f. (6)) is twice the difference of two sigmoid functions (which is bounded between 1 and -1), multiplied by the covariate vector x which by assumption has norm at most B. Thus, either $\bar{\ell}(\theta; \theta_*) > \bar{\ell}(\theta_*; \theta_*) - \varepsilon$, for some $\varepsilon > 0$, or else $(\theta^\top x - \theta_*^\top x)^2 \geq (\frac{\varepsilon}{2B})^2$. Making use of this and also bounding $\left(\sigma\left(a - \theta^\top x\right) - \sigma\left(b - \theta^\top x\right)\right)$ via α (as in fact the former is the probability of observing the sample x, and the latter is the same probability conditioned on a specific class):

$$\begin{aligned}\langle \nabla_\theta \bar{\ell}(\theta; \theta_*), \theta - \theta_* \rangle &< C_0 \cdot \left(\sigma\left(a - \theta^\top x\right) - \sigma\left(b - \theta^\top x\right)\right) \cdot (\theta^\top x - \theta_*^\top x)^2 \\ &< -\frac{C_0 \cdot \alpha \cdot \varepsilon^2}{4 \cdot B^2}. \\ &= -\frac{\alpha^2 \cdot \varepsilon^2}{4 \cdot B^2} \cdot \frac{(1-e^a)(e^b-1)}{e^b - e^a}.\end{aligned}$$

This concludes the proof. □

C.4 Quasiconvexity of empirical log-likelihood via concentration

Lemma 4. *For a minibatch size $b \geq \bar{\Theta}\left(poly(B, \frac{1}{\alpha}) \cdot \frac{1}{\varepsilon^2}\right)$, the minibatch log-likelihood, $\ell_b(\theta;\theta_*) = \sum_{i=1}^{b} \ell(\theta; x_i, y_i)$, is $(\varepsilon, \kappa, \theta_*)$-SLQC where $\kappa \in \Theta\left(\frac{1}{\epsilon} \cdot poly\left(B, \frac{1}{\alpha}\right)\right)$.*

Proof. Note that in this proof we carry forward our notation, and in particular see Appendix C.3 for the definition of C_0. Our approach is to use concentration of measure to show that for large enough batch size, the empirical log-likelihood exhibits quasi-convexity properties similar to those of the population log-likelihood. We can use α to bound the difference $\theta^\top x - \theta_*^\top x$ as follows:

$$\langle \nabla_\theta \widehat{\ell}(\theta, x, y), \theta - \theta_* \rangle \leq 2 \cdot \log\left(\frac{1}{\alpha} - 1\right) \leq 2 \cdot \log(1/\alpha).$$

Also note that by definition for any fixed θ,

$$\sum_{x_i \in \mathcal{X}} \mathbb{E}_{y_i}\left[\langle \nabla_\theta \widehat{\ell}(\theta, x, y), \theta - \theta_* \rangle\right] = \langle \nabla_\theta \bar{\ell}(\theta; \theta_*), \theta - \theta_* \rangle.$$

To this end, we define K_n as the inner product $\langle \sum_{i=1}^{n} \nabla_\theta \widehat{\ell}(\theta, x_i, y_i), \theta - \theta_* \rangle$ for samples $(x_i \in \mathcal{X}, y_i)$. We can thus apply Hoeffding's inequality as follows:

$$\mathbb{P}\left(K_n \geq \langle \nabla_\theta \bar{\ell}(\theta; \theta_*), \theta - \theta_* \rangle + t\right) \leq \exp\left\{-\frac{nt^2}{2\log^2(1/\alpha)}\right\}$$

Based on the result of Lemma 3 in the last section, we set $t = \frac{\varepsilon^2 \cdot \alpha \cdot C_0}{8 \cdot B^2}$, we set $n \geq \frac{64 \cdot B^4 \cdot \log^2(1/\alpha)}{\alpha^2 \cdot C_0^2 \cdot \varepsilon^2} \cdot \delta$ such that

$$\mathbb{P}\left(K_n \geq -\frac{\epsilon^2 \cdot \alpha \cdot C_0}{8 \cdot B^2}.\right) \leq \exp\{-\delta\}$$

Now, to prove that the log-likelihood is locally quasiconvex, we once again use the fact that the likelihood is $2B$-Lipschitz, and thus $\langle \nabla_\theta \bar{\ell}(\theta, \theta_*), \theta - \widehat{\theta} \rangle \leq 2B \cdot \|\theta - \widehat{\theta}\|_2$. Thus, for any $\widehat{\theta}$ satisfying

$$\|\widehat{\theta} - \theta\|_2 \leq \frac{\epsilon^2 \cdot \alpha \cdot C_0}{16 \cdot B^3},$$

we have that

$$\langle \nabla_\theta \bar{\ell}(\theta, \theta_*), \widehat{\theta} - \theta_* \rangle \leq 0 \qquad \text{w.p. } \exp\{-\delta\}.$$

We next bound C_0 as follows:

$$\begin{aligned}
C_0 = \frac{\alpha}{2} \cdot \frac{(1 - e^a)(e^b - 1)}{e^b - e^a} &= \frac{\alpha}{2} \cdot (1 - e^a)(1 - e^{-b}) \\
\text{By definition, } \sigma(-a) \geq \alpha + \frac{1}{2} \qquad &\implies \qquad -a \geq \log\left(\frac{\alpha + \frac{1}{2}}{\frac{1}{2} - \alpha}\right) \\
e^a \leq \frac{\frac{1}{2} - \alpha}{\frac{1}{2} + \alpha} \qquad &\implies \qquad 1 - e^a \geq \frac{2\alpha}{\frac{1}{2} + \alpha} \\
\text{By symmetry, } C_0 &\geq \frac{2 \cdot \alpha^3}{(\frac{1}{2} + \alpha)^2} \\
&\geq \frac{8}{9} \cdot \alpha^3.
\end{aligned}$$

This implies that $\ell_b(\theta, \theta_*)$ with $b \geq \bar{\Theta}\left(\text{poly}(\frac{1}{\alpha}, B), \frac{1}{\varepsilon^2}\right) \cdot \delta$ samples is $(\varepsilon, \frac{9 \cdot B^3}{4 \cdot \varepsilon \cdot \alpha^4}, \theta_*)$-SLQC in θ so long as $\theta \in R_\alpha$.

□

D Detailed Experimental Setup

D.1 Synthetic Data

The data generation model: We start by computing $X_0 = \{x_1, \ldots, x_n : \; x_i \sim \mathcal{U}\left([0, 100]\right)^2\}$, an 10000×2 randomly generated data matrix. A "ground-truth" θ_* is selected at random ($\theta_* \sim \mathcal{U}([-1, 1])$), then for each x_i, a latent z_i is sampled according to $z_i = \theta_*^\top x_i + \varepsilon$, where ε is a standard Gaussian (Logistic) random variable in the case of probit (logistic) regression. Finally, we remove all (x, z) pairs where $z \leq C$ for some predetermined threshold C (truncation) and label each x_i with $y_i = \mathbf{1}_{z_i \geq 0}$.

Methodology: In both cases, we use gradient descent with 1000 iterations in order to find the estimated parameter—in all cases the training loss plateaus. Since the data was generated according to a ground-truth θ_*, we evaluate each method based on cosine similarity between the estimated parameter θ and the true parameter θ_*.

D.2 Neural Network: Rotation Prediction

Here we give further detail on the experimental setup for the rotation prediction experiment in Section 4. In this experiment, we sample images x uniformly from the CIFAR-10 training set. A random rotation of $d = \mathcal{U}([0, 360])$ degrees is then applied to the image, and the corresponding label is $z = d + \mathcal{N}(0, \sigma)$ for some value of σ. We then discard all of the points for which $z \notin [0, 180]$, and train on the remaining points with the hyperparameters given in Table 1. We train both minimizing standard squared loss between the predicted angle and the true angle, as well as a custom implemented loss whose gradient is implemented to be the analog for the truncated log-likelihood case.

D.3 Neural Network: Class Truncation

Here we give further detail on the experimental setup for the CNN classification experiment. We begin by training a base model on the binary task of classifying CIFAR-10 dogs vs. cats (with the cross-entropy loss and the hyperparameters given in Table 1. Let h_θ represent the function mapping images to the log-probability assigned to "dog" by the base model (i.e. the output of the model pre-sigmoid). Then, for each image x_i, we calculate $z_i = h_\theta(x_i) + \mathcal{N}(0, 1)$. For some varying truncation parameter C, we remove all the samples with $z_i < C$, and label the remaining samples with $y_i = \mathbf{1}_{z_i \geq 0}$. We then train models using both the standard cross-entropy loss and a custom version implementing the truncated likelihood gradient. We vary the truncation parameter C and train the networks with the hyperparameters given in Table 1. We then test on the original dogs vs. cats task.

	Rotation	Classification
Learning Rate	5e-6	0.1
Weight decay	5e-4	5e-4
Epochs	20	150
LR drop frequency	8	50
Momentum	0.9	0.9
Data augmentation	Y	Y

Table 1: Hyperparameters for neural network experiments.

E Code for Regression and Classification

```python
PREDICATE = ... # boolean function

class TruncatedMSE(ch.autograd.Function):
    @staticmethod
    def forward(ctx, pred, targ):
        ctx.save_for_backward(pred, targ)
        return 0.5 * (pred.float() - targ.float()).pow(2).mean()

    @staticmethod
    def backward(ctx, grad_output):
        pred, targ = ctx.saved_tensors
        # Make args.num_samples copies of pred, N x B x 1
        stacked = pred[None,...].repeat(args.num_samples,1,1)
        # Add random noise to each copy
        noised = stacked + ch.randn_like(stacked)
        # Filter out the copies where pred is in bounds
        filtered = PREDICATE(noised).float()
        # Average over truncated indices
        out = (noised * filtered).sum(dim=0) / (filtered.sum(dim=0) + args.eps)
        grad = ch.where(out > 0, out, targ) - targ
        return grad / pred.shape[0], \
                 (targ - pred) / pred.shape[0]
```

Listing 1: Truncated version of the mean squared-error loss

```python
import torch as ch
from torch.nn import functional as F
from torch import sigmoid as sig
from torch.distributions import transforms, Uniform, TransformedDistribution

base_distribution = Uniform(0, 1)
transforms = [transforms.SigmoidTransform().inv]
logistic = TransformedDistribution(base_distribution, transforms)

class TruncatedBCE(ch.autograd.Function):
    @staticmethod
    def forward(ctx, pred, targ):
        pred, targ = pred, targ.float()
        ctx.save_for_backward(pred, targ)
        return F.binary_cross_entropy_with_logits(pred, targ)

    @staticmethod
    def backward(ctx, grad_output):
        pred, targ = ctx.saved_tensors
        stacked = pred[None,...].repeat(args.num_samples,1,1)
        noised = stacked + logistic.sample()
        filtered = (noised > args.C).float()
        out = (noised * filtered).sum(dim=0) / (filtered.sum(dim=0) + 1e-5)
        grad = ch.where(ch.abs(out) > 1e-5, sig(out), targ) - targ

        N = pred.shape[0]
        return grad / N, -grad / N
```

Listing 2: Truncated version of the binary cross-entropy loss