The Annals of Statistics
2014, Vol. 42, No. 5, 1693–1724
DOI: 10.1214/14-AOS1220

LOCAL CASE-CONTROL SAMPLING: EFFICIENT SUBSAMPLING IN IMBALANCED DATA SETS

BY WILLIAM FITHIAN[1] AND TREVOR HASTIE[2]

Stanford University

For classification problems with significant class imbalance, subsampling can reduce computational costs at the price of inflated variance in estimating model parameters. We propose a method for subsampling efficiently for logistic regression by adjusting the class balance locally in feature space via an accept–reject scheme. Our method generalizes standard case-control sampling, using a pilot estimate to preferentially select examples whose responses are conditionally rare given their features. The biased subsampling is corrected by a post-hoc analytic adjustment to the parameters. The method is simple and requires one parallelizable scan over the full data set.

Standard case-control sampling is inconsistent under model misspecification for the population risk-minimizing coefficients θ^*. By contrast, our estimator is consistent for θ^* provided that the pilot estimate is. Moreover, under correct specification and with a consistent, independent pilot estimate, our estimator has exactly twice the asymptotic variance of the full-sample MLE—even if the selected subsample comprises a miniscule fraction of the full data set, as happens when the original data are severely imbalanced. The factor of two improves to $1 + \frac{1}{c}$ if we multiply the baseline acceptance probabilities by $c > 1$ (and weight points with acceptance probability greater than 1), taking roughly $\frac{1+c}{2}$ times as many data points into the subsample. Experiments on simulated and real data show that our method can substantially outperform standard case-control subsampling.

1. Introduction. In recent years, statisticians, scientists and engineers are increasingly analyzing enormous data sets. When data sets grow sufficiently large, computational costs may play a major role in the analysis, potentially constraining our choice of methodology or the number of data points we can afford to process. Computational savings can translate directly to statistical gains if they:

(1) enable us to experiment with and prototype a variety of models, instead of trying only one or two,

(2) allow us to refit our models more often to adapt to changing conditions,

(3) allow for cross-validation, bagging, boosting, bootstrapping or other computationally intensive statistical procedures or

Received June 2013; revised March 2014.

[1]Supported by NSF VIGRE Grant DMS-05-02385.

[2]Supported in part by NSF Grant DMS-10-07719 and NIH Grant RO1-EB001988-15.

MSC2010 subject classifications. Primary 62F10; secondary 62D05.

Key words and phrases. Logistic regression, case-control sampling, subsampling.

(4) open the door to using more sophisticated statistical techniques on a compressed data set.

Bottou and Bousquet (2008) discuss the tradeoffs arising when we adopt this point of view. One simple manifestation of these tradeoffs is that we may run out of computing resources before we run out of data, in effect making the sample size n a function of the efficiency of our fitting method.

1.1. *Imbalanced data sets*. Class imbalance is pervasive in modern classification problems and has received a great deal of attention in the machine learning literature [Chawla, Japkowicz and Kotcz (2004)]. It can come in two forms:

Marginal imbalance. One of the classes is quite rare; for instance, $\mathbb{P}(Y = 1) \approx 0$. Such imbalance typically occurs in data sets for predicting click-through rates in online advertising, detecting fraud or diagnosing rare diseases.

Conditional imbalance. For most values of the features X, the response Y is very easy to predict; for instance, $\mathbb{P}(Y = 1|X = 0) \approx 0$ but $\mathbb{P}(Y = 1|X = 1) \approx 1$. For example, such imbalance might arise in the context of email spam filtering, where well-trained classifiers typically make very few mistakes.

Both or neither of the above may occur in any given data set. The machine learning literature on class imbalance usually focuses on the first type, but the second type is also common.

If, for example, our data set contains one thousand or one million negative examples for each positive example, then many of the negative data points are in some sense redundant. Typically in such problems, the statistical noise is primarily driven by the number of representatives of the rare class, whereas the total size of the sample determines the computational cost. If so, we might hope to finesse our computational constraints by subsampling the original data set in a way that enriches for the rare class. Such a strategy must be implemented with care if our ultimate inferences are to be valid for the full data set.

This article proposes one such data reduction scheme, local case-control sampling, for use in fitting logistic regression models. The method requires one parallelizable scan over the full data set and yields a potentially much smaller subsample containing roughly half of the information found in the original data set.

1.2. *Subsampling*. The simplest way to reduce the computational cost of a procedure is to subsample the data before doing anything else. However, uniform subsampling from an imbalanced data set is inefficient since it fails to exploit the unequal importance of the data points.

Case-control sampling—sampling uniformly from each class but adjusting the mixture of the classes—is a more promising approach. This procedure originated in epidemiology, where the positive examples (cases) are typically diseased patients and negative examples (controls) are disease-free [Mantel and Haenszel

(1959)]. Often, an equal number of cases and controls are sampled, resulting in a subsample with no marginal imbalance, and costly measurements of predictor variables are only made for selected patients [Breslow, Day et al. (1980)]. This method is useful in our context as well, since a logistic regression model fitted on the subsample can be converted to a valid model for the original population via a simple adjustment to the intercept [Anderson (1972), Prentice and Pyke (1979)].

However, standard case-control sampling still may not make most efficient use of the data. For instance, it does nothing to exploit conditional imbalance in a data set that is marginally balanced. Even with some marginal imbalance, a control that looks similar to the cases is often more useful for discrimination purposes than one that is obviously not a case.

We propose a method, local case-control sampling, which attempts to remedy imbalance *locally* throughout the feature space. Given a pilot estimate $(\tilde{\alpha}, \tilde{\beta})$ of the logistic regression parameters, local case-control sampling preferentially keeps data points for which Y is surprising given X. Specifically, if $\tilde{p}(x) = \frac{e^{\tilde{\alpha}+\tilde{\beta}'x}}{1+e^{\tilde{\alpha}+\tilde{\beta}'x}}$, we accept (x_i, y_i) with probability $|y_i - \tilde{p}(x_i)|$, the ℓ_1 residual of the pilot model. In the presence of extreme marginal or conditional imbalance, these errors will generally be quite small and the subsample can be many orders of magnitude smaller than the full data set.

Just as with case-control sampling, we can fit our model to the subsample and make an equally simple correction to obtain an estimate for the original data set. When the logistic regression model is correctly specified and the pilot is consistent and independent of the data, the asymptotic variance of the local case-control estimate is exactly twice the variance of a logistic regression fit on the (potentially much larger) full data set. This factor of two improves to $1 + \frac{1}{c}$ if we accept with probability $c|y_i - \tilde{p}(x_i)| \wedge 1$ and weight accepted points by a factor of $c|y_i - \tilde{p}(x_i)| \vee 1$. For example, if $c = 5$ then the variance of the subsampled estimate is only 20% greater than the variance of the full-sample MLE. The subsample we take with $c > 1$ is no more than c times larger than the subsample for $c = 1$, and for data sets with large imbalance is roughly $\frac{1+c}{2}$ times as large.

Local case-control sampling also improves on the bias of standard case-control sampling. When the logistic regression model is misspecified, case-control sampling is in general inconsistent for the risk minimizer in the original population. By contrast, local case-control sampling is always consistent given a consistent pilot, and is also asymptotically unbiased when the pilot is. Sections 5 and 6 present empirical results demonstrating the advantages of our approach in simulations and on the Yahoo! webspam data set.

1.3. *Notation and problem setting.* Our setting is that of predictive classification: we are given n independent and identically distributed observations, each consisting of predictors $x_i \in \mathcal{X}$ and a binary response $y_i \in \{0, 1\}$, with joint probability measure $\mathbb{P}$. For our purposes, we assume the predictors are mapped into

some real covariate vector space, so that $\mathcal{X} \subseteq \mathbb{R}^p$ Our aim is to learn the function

$$p(x) = \mathbb{P}(Y = 1 | X = x) \tag{1}$$

or equivalently to learn

$$f(x) = \operatorname{logit}\big(p(x)\big) = \log \frac{p(x)}{1 - p(x)} \tag{2}$$

which could be infinite for some x.

A linear logistic regression model assumes f is linear in x; that is,

$$f_\theta(x) = f_{\alpha,\beta}(x) = \alpha + \beta' x, \tag{3}$$

where $\theta = (\alpha, \beta) \in \mathbb{R}^{p+1}$. This is less of a restriction than it might seem, since x may represent a very large basis expansion of some smaller set of "raw" features.

Nevertheless, in the real world, f is unlikely to satisfy our parametric model for any given basis x. When the model is misspecified, we can still view logistic regression as an M-estimator with convex loss equal to the negative log-likelihood for a single data set:

$$\rho(\theta; x, y) = -y\big(\alpha + \beta' x\big) + \log\big(1 + e^{\alpha + \beta' x}\big). \tag{4}$$

As an M-estimator, under general conditions logistic regression in large samples will converge to the minimizer of the population risk $R(\theta) = \mathbb{E}\rho(\theta; X, Y)$ [Huber (2011)]. That is, θ converges to the population maximizer of the expected log-likelihood

$$\theta^* = \arg\min_\theta \mathbb{E}\rho(\theta; X, Y) \tag{5}$$

$$= \arg\min_\theta \mathbb{E}\big[-Y\big(\alpha + \beta' X\big) + \log\big(1 + e^{\alpha + \beta' X}\big)\big]. \tag{6}$$

If $f = f_{\theta_0}$ for some θ_0, then $\theta^* = \theta_0$; otherwise f_{θ^*} is the best linear approximation to f in the sense of (5). For a misspecified model, $f_{\hat\theta}$ cannot possibly converge to f no matter what sampling scheme or estimation procedure we use, or how much data we obtain. Consistency, then, will mean that $\hat\theta \xrightarrow{p} \theta^*$.

Model misspecification is ubiquitous in real-world applications of regression methods. For reasons of exposition, the misspecification always takes a simple form in our simulations, for example, in Example 1 there are two binary predictors, and we would have correct specification if only we added one interaction—but in the real world it usually is neither possible nor even desirable to expand the feature basis until the model is correctly specified. For instance, if $p = 1000$, then there are $\binom{p+1}{2} = 500{,}500$ quadratic terms. Even if we included all those terms as features, we would still be missing cubic terms, quartic terms, and so on.

Some kinds of misspecification are milder than others, and some are easier to find and fix than others. Seeking better-specified models (without adding too

much model complexity) is a worthy goal, but realistically perfect specification is unattainable.

Our goal, then, is to speed up computation while still obtaining a good estimate of θ^*, the population logistic regression parameters. As we will see, standard case-control sampling achieves the first goal, but may fail at the second.

1.4. *Related work.* Recent years have seen substantial work on classification in imbalanced data sets. See Chawla, Japkowicz and Kotcz (2004) and He and Garcia (2009) for surveys of machine learning efforts on this problem. Many of the methods proposed involve some form of undersampling the majority class, oversampling the minority class, or both. Owen (2007) examined the limit of marginally imbalanced logistic regression and proved it is equivalent to fitting an exponential family model to the minority class.

One recurring theme is to preferentially sample negative examples that lie near positive examples in feature space. For example, Mani and Zhang (2003) propose selecting majority-class examples whose average distance to its three nearest minority examples is smallest. Our method has a similar flavor since the probability of sampling a negative example $(x, 0)$ is $\tilde{p}(x)$, which is large when the features x are similar to those characteristic of positive examples.

Our proposal lies more in the tradition of the epidemiological case-control sampling literature. In particular, case-control sampling within several categorical strata has been studied by Breslow and Cain (1988), Fears and Brown (1986), Scott and Wild (1991), Weinberg and Wacholder (1990). Typically, the strata are based on easy-to-measure screening variables available for a wide population, with more laborious-to-collect variables being measured on the sampled subjects. Lumley, Shaw and Dai (2011) discuss survey calibration methods for efficient regression in two-stage sampling schemes, which are interesting but too computationally intensive for our purposes here.

2. Case-control subsampling. Case-control sampling is commonly carried out by taking all the cases and exactly c times as many controls for some fixed c (e.g., $c = 1, 2, 5$). However, for our purposes it will be simpler to consider a nearly equivalent procedure based on accept–reject sampling.

Define some acceptance probability function $a(y)$ and let $b = \log \frac{a(1)}{a(0)}$, the log-selection bias. Consider the following algorithm:

(1) Generate independent $z_i \sim \text{Bernoulli}(a(y_i))$.

(2) Fit a logistic regression to the subsample $S = \{(x_i, y_i) : z_i = 1\}$, obtaining unadjusted estimates $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S)$.

(3) Assign $\hat{\alpha} \leftarrow \hat{\alpha}_S - b$ and $\hat{\beta} \leftarrow \hat{\beta}_S$.

Specifically, we could generate the z_i by first generating $u_i \sim U(0, 1)$ mutually independent of the pilot, the data, and each other, then taking $z_i = \mathbf{1}_{u_i \le a(y_i)}$. Note

that steps (2)–(3) are equivalent to logistic regression with offset b for each data point.

This variant is convenient to analyze because the subsample thus obtained is an i.i.d. sample from a new population:

$$\mathbb{P}_S(X, Y) = \mathbb{P}(X, Y|Z = 1) = \frac{a(Y)\mathbb{P}(X, Y)}{\bar{a}} \tag{7}$$

with $\bar{a} = a(1)\mathbb{P}(Y = 1) + a(0)\mathbb{P}(Y = 0)$, the marginal probability of $Z = 1$.

The estimate $(\hat{\alpha}, \hat{\beta})$ is motivated by a simple application of Bayes' rule relating the odds of $Y = 1$ in $\mathbb{P}$ and $\mathbb{P}_S$. If $g(x)$ is the true conditional log-odds function for $\mathbb{P}_S$, we have

$$g(x) = \log \frac{\mathbb{P}(Y = 1|X = x, Z = 1)}{\mathbb{P}(Y = 0|X = x, Z = 1)} \tag{8}$$

$$= \log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} + \log \frac{\mathbb{P}(Z = 1|Y = 1, X = x)}{\mathbb{P}(Z = 1|Y = 0, X = x)} \tag{9}$$

$$= f(x) + b. \tag{10}$$

That is, the log-odds $g(x)$ in our biased population is simply a vertical shift by b of the log-odds $f(x)$ in the original population, so given an estimate of g we can subtract b to estimate f. If the model is correctly specified, logistic regression on the subsample yields a consistent estimate for the function $g(x)$, so the estimate for $f(x)$ is also consistent.

Note that the derivation (8)–(10) is equally valid if the sampling bias b depends on x, in which case we have $g(x) = f(x) + b(x)$. Local case-control sampling exploits this more general identity.

2.1. *Conditional probability and the logit loss.* If X is integrable, then upon differentiating the population risk (5) with respect to θ we obtain the population score criterion:

$$0 = \mathbb{E}\left[\left(Y - \frac{e^{f_\theta(X)}}{1 + e^{f_\theta(X)}}\right)\begin{pmatrix}1\\ X\end{pmatrix}\right] = \int \left(p(x) - p_\theta(x)\right)\begin{pmatrix}1\\ x\end{pmatrix} d\mathbb{P}(x). \tag{11}$$

Informally, the best linear predictor is the one that gets the conditional probabilities right on average. Note this is not the same as a predictor that gets the conditional log-odds right on average.

To illustrate the difference between approximating probabilities and approximating logits, suppose that $X \sim U(0, 1)$ and $f(x) = -10 + 5x + 3 \cdot \mathbf{1}_{x>0.5}$. The left panel of Figure 1 shows $f(x)$ as a solid line and its best linear approximation as a dashed line. On the logit scale, the dashed line appears to be a very poor fit to the black curve. It fits reasonably well for large x, but it appears more or less to ignore the smaller values of x.

The right panel of Figure 1 shows why. When we transform both curves to the probability scale, the fit looks much more reasonable. $f_{\theta^*}(x)$ need not approximate

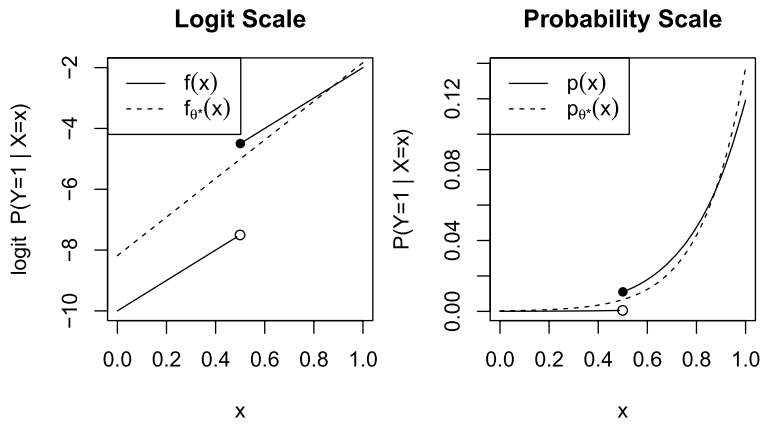

FIG. 1. *The best linear fit $f_{\theta^*}(x)$ approximates the true log-odds $f(x)$ in the sense of matching its implied conditional probabilities, not logits.*

$f(x)$ particularly well for small x, because in that range even a large change in the log-odds produces a negligible change in the conditional probability $p(x)$. By contrast, $f_{\theta^*}(x)$ needs to approximate $f(x)$ well for larger x, where $p(x)$ changes more rapidly.

In general, logistic regression places highest priority on fitting f where $\frac{dp(x)}{df(x)}$ is largest: where $f(x) \approx 0$ and $p(x) \approx 0.5$. In this example, with its strong marginal imbalance, the regions that matter most are those where $p(x)$ is largest. This often makes sense in applications such as medical screening or advertising click-through rate prediction, where accuracy is most important when the probability of disease or click-through is nonnegligible. In Section 7, we consider how to modify the method to obtain classifiers that prioritize correctness near some other, user-defined level curve of $p(x)$.

Finally, note that Figure 1 suggests the case-control sampling estimate is unlikely to be consistent for θ^* in general. The nature of our linear approximation in the left panel is intimately related to the fact that $f(x) < 0$ everywhere in the sample space. If $f(x)$ were shifted upward by some constant, the response of the dashed curve would be more complicated than a simple constant shift by b, since the relative importance of the two segments would change. Therefore, estimating $f(x) + b$ and then subtracting b may not be a successful strategy.

2.2. *Inconsistency of case-control under misspecification.* If the linear model is misspecified, the case-control estimate is generically not consistent for the best linear predictor θ^* as $n \to \infty$ [Manski and Thompson (1989), Xie and Manski (1989)]. The unadjusted estimate will instead converge to the best linear predictor of g for the distribution $\mathbb{P}_S$, which solves the score criterion

$$0 = \int \left(\frac{e^{f(x)+b}}{1+e^{f(x)+b}} - \frac{e^{f_\theta(x)}}{1+e^{f_\theta(x)}} \right) \begin{pmatrix} 1 \\ x \end{pmatrix} d\mathbb{P}_S(x). \tag{12}$$

Let $\theta^*_{\mathrm{CC}}(b)$ be the large-sample limit of the *adjusted* case-control sampling estimate with bias b. Then $\theta^*_{\mathrm{CC}}(b)$ solves the population score criterion

$$0 = \int \left(\frac{e^{f(x)+b}}{1+e^{f(x)+b}} - \frac{e^{f_\theta(x)+b}}{1+e^{f_\theta(x)+b}} \right) \begin{pmatrix} 1 \\ x \end{pmatrix} d\mathbb{P}_S(x) \tag{13}$$

which differs from (11) in two ways. First, the integral is taken over a different distribution for X. Second, and more importantly, the integrand is different. We are now approximating $f(x)$ in a different sense than we were.

In general under misspecification, $\theta^*_{\mathrm{CC}}(b)$ is different for every b. If we sample cases and controls equally, in the limit we will get a different answer than if we sample twice as many controls; and in either case we will get a different answer than if we use the entire data set or subsample uniformly.

These differences can be quite consequential for our inferences about β or the predictive performance of our model, as we see next.

EXAMPLE 1 (Oatmeal and disease risk). In this fictitious example, we consider estimating the effect of exposure to oatmeal on a person's risk of developing some rare disease. Suppose that 10% of the population has a family history of the disease, half the population eats oatmeal (independently of family history), and that both exposure and family history are binary predictors. Suppose further that the true conditional log-odds function $f(x)$ is given by the top-left panel of Table 1.

The corresponding conditional probabilities $p(x)$ are given in the lower-left panel of Table 1. Notice that oatmeal increases the risk for people who are already at risk by virtue of their family history, but has a protective effect for everyone else. This interaction means that the additive logistic regression model is misspecified.

TABLE 1
Disease risk in the full population, and in the population created by case-control sampling with equal numbers in each class

Original population ($\mathbb{P}$)

Conditional log-odds (f)

	History −	**History +**
Oatmeal −	−5	−4
Oatmeal +	−10	−1

Case-control population ($\mathbb{P}_S$)

Conditional log-odds (g)

	History −	**History +**
Oatmeal −	−1.2	−0.2
Oatmeal +	−6.2	2.8

Conditional probabilities

	History −	**History +**
Oatmeal −	0.007	0.02
Oatmeal +	5E−5	0.37

Conditional probabilities

	History −	**History +**
Oatmeal −	0.24	0.46
Oatmeal +	0.002	0.94

Because only the probabilities in the "History +" column are large enough to matter, the fitted model for $f(x)$ pays more attention to the at-risk population, for whom oatmeal elevates the risk of disease. A logistic regression on a large sample from this population estimates the coefficient for oatmeal as $\beta^*_{\text{Oatmeal}} = 1.4$, implying an odds ratio of about 4.0. This is close to the marginal odds ratio of roughly 4.3 that we would obtain if we did not control for family history.

Suppose, however, that we sampled an equal number of cases and controls. Then the conditional log-odds of disease in our sample would reflect the top-right panel of Table 1, with all cells increased by the same amount.

For large samples, the case-control estimate is $\beta^*_{\text{CC,Oatmeal}} = -0.83$, implying an odds ratio of about 0.44. Using case-control sampling has reversed our inference about the effect of oatmeal exposure, because after shifting the log-odds the left column becomes much more important.

EXAMPLE 2 (Two-class Gaussian model). Suppose that $\mathbb{P}(Y = 1) = 1\%$, and that $X|Y \sim \mathrm{N}(\mu_Y, \Sigma_Y)$. Let

$$\mu_0 = (0, 0), \qquad \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{14}$$

$$\mu_1 = (1.5, 1.5), \qquad \Sigma_1 = \begin{pmatrix} 0.3 & 0 \\ 0 & 5 \end{pmatrix}. \tag{15}$$

Data simulated from this model are shown in the left panel of Figure 2. In this example, the true log-odds $f(x)$ is an additive quadratic function of the two coordinates X_1 and X_2.

In this example as in the previous one, the population-optimal case-control parameters differ substantially from the optimal parameters in the original population, with dramatic effects for the predictive performance of the model. The decision boundaries for the two estimates are overlayed on the left panel of Figure 2. In the right panel, we plot the precision–recall curves resulting from each set of parameters on a large test set.

2.3. *Weighted case-control sampling.* A simple alternative to standard case-control sampling is to weight the subsampled data points by the inverse of their probability of being sampled. We include weighted case-control sampling as a competitor in our simulation studies in Section 5. Because it is a Horvitz–Thompson estimator with positive sampling probabilities for any (x, y) pair, this method is $\sqrt{n}$-consistent, and asymptotically normal and unbiased under general conditions [Horvitz and Thompson (1952)].

Although weighting succeeds in removing the bias induced by the case-control sampling, this consistency comes at a cost of increasing the variance, since the effective sample size is reduced [Scott and Wild (1986, 2002)].

Despite its inefficiency, the weighted case-control method can be an attractive means of obtaining a consistent pilot if another good pilot is not immediately available, and we later will use it to that end in our experiments.

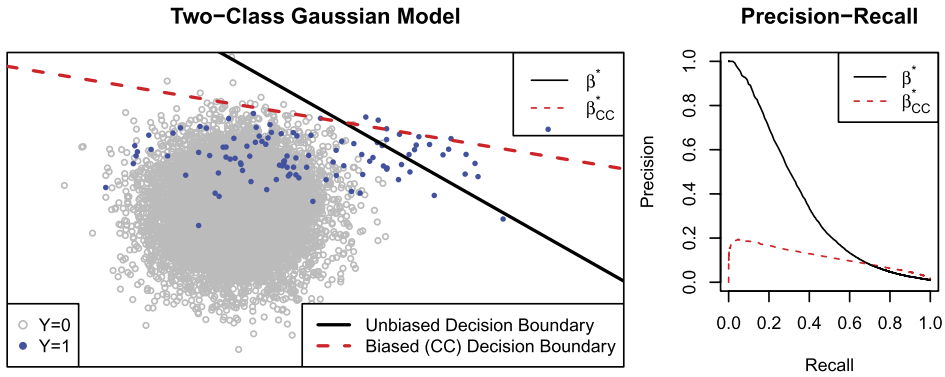

FIG. 2. *At left, biased (case-control) and unbiased decision boundaries for the bivariate Gaussian mixture model. At right, precision–recall curves for* β^* *and* β^*_{CC}.

3. Local case-control subsampling. In this section, we describe local case-control subsampling, a generalization of standard case-control sampling that both improves on its efficiency and resolves its problem of inconsistency. To achieve these benefits, we require a pilot estimate, that is, a good guess $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta})$ for the population-optimal θ^*.

3.1. *The local case-control sampling algorithm.* Local case-control sampling differs from case-control sampling only in that the acceptance probability a is allowed to depend on x as well as y. Our criterion for selection will be the degree of "surprise" we experience upon observing y_i given x_i:

$$a(x, y) = |y - \tilde{p}(x)| = \begin{cases} 1 - \tilde{p}(x), & y = 1, \\ \tilde{p}(x), & y = 0, \end{cases} \tag{16}$$

where $\tilde{p}(x) = \frac{e^{\tilde{\alpha}+\tilde{\beta}'x}}{1+e^{\tilde{\alpha}+\tilde{\beta}'x}}$ is the pilot estimate of $\mathbb{P}(Y = 1|X = x)$. The algorithm is:

(1) Generate independent $z_i \sim \text{Bernoulli}(a(x_i, y_i))$.
(2) Fit a logistic regression to the sample $S = \{(x_i, y_i) : z_i = 1\}$ to obtain unadjusted estimates $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S)$.
(3) Assign $\hat{\alpha} \leftarrow \hat{\alpha}_S + \tilde{\alpha}$ and $\hat{\beta} \leftarrow \hat{\beta}_S + \tilde{\beta}$.

As before, steps (2)–(3) are equivalent to fitting a logistic regression in the subsample with offsets $-\tilde{\alpha} - \tilde{\beta}'x_i$. The z_i are generated as in Section 2, and the adjustment is again justified by (8)–(10), only now with the constant b replaced by

$$b(x) = \log\left(\frac{a(x, 1)}{a(x, 0)}\right) = -\tilde{\alpha} - \tilde{\beta}'x. \tag{17}$$

In other words, the subsample is drawn from a measure with

$$g(x) = f(x) - \tilde{\alpha} - \tilde{\beta}'x. \tag{18}$$

If $f(x)$ is well approximated by the pilot estimate, then $g(x) \approx 0$ throughout feature space. That is, conditional on selection into S, y_i given x_i is nearly a fair coin toss.

To motivate this choice heuristically, recall that the Fisher information for the log-odds of a Bernoulli random variable is maximized when the probability is $\frac{1}{2}$: fair coin tosses are more informative than heavily biased ones. In effect, local case-control sampling tilts the conditional distribution of Y given $X = x$ to make each y_i in the subsample more informative. We then fit a logistic regression in the more favorable sampling measure, and "tilt back" to obtain a valid estimate for the original population.

In marginally imbalanced data sets where $\mathbb{P}(Y = 1|X = x)$ is small everywhere in the predictor space, a good pilot has $\tilde{p}(x) \approx 0$ for all x, and the number of cases discarded by this algorithm will be quite small. If we wish to avoid discarding any cases, we can always modify the algorithm so that instead of keeping $(x, 1)$ with probability $a(x, 1)$, we keep it with probability 1 and assign weight $a(x, 1)$.

3.2. *Choosing the pilot fit.* In many applications, there may be a natural choice of pilot fit $\tilde{\theta}$; for instance, if we are refitting a classification model every day to adapt to a changing world, then yesterday's fit is a natural choice for today's pilot.

If no pilot fit is available from such a source, we recommend an initial pass of weighted case-control sampling (described in Section 2.3) to obtain the pilot. Weighted case-control sampling using a fixed fraction of the original sample is itself $\sqrt{n}$-consistent and asymptotically unbiased for the true parameters. Consequently, if the pilot were fit using an independent data set the second-stage estimate would enjoy consistency and asymptotic unbiasedness per the results in Section 4.

Our experiments suggest that mild inaccuracy in the pilot estimate, and using a data-dependent pilot, do not unduly degrade the performance of the local case-control algorithm. For example, is Simulation 2 of Section 5.2, the pilot is fifty times less efficient than the final local case-control estimate. The main role of the pilot fit is to guide us in discarding most of the data points for which y_i is obvious given x_i while keeping those for which y_i is conditionally surprising.

In our example and simulations, we use a pilot sample about the same size as the local case-control subsample, on the principle that we can afford to spend about as much time computing the pilot as computing the second-stage estimate. When $\mathbb{P}(Y = 1|X)$ is small throughout $\mathcal{X}$, this rule amounts roughly to weighted case-control sampling using all the cases and one control per case. Although the above rule has worked reasonably well for us, at this time we can offer no finite-sample guarantees that any given pilot sample size is large enough.

Because standard case-control sampling amounts to local case-control sampling with a constant-only pilot fit, we might expect that the pilot fit need not be perfect to improve upon case-control sampling. Our experiments in Sections 5 and 6 support this intuition.

3.3. *Taking a larger or smaller sample.* As we will see in Section 4.3, under correct model specification, and with an independent and consistent pilot, the baseline procedure described above has exactly twice the asymptotic variance as a logistic regression estimated with the full sample, despite using a potentially very small subset of the data. We can improve upon this factor of two by increasing the size of the subsample.

One simple way to achieve this is to multiply all acceptance probabilities by some constant c, for example, $c = 5$. When deciding whether to sample the point (x_i, y_i), we would then generate $z_i \sim \text{Bernoulli}(ca(x_i, y_i) \wedge 1)$ and assign weight $w_i = ca(x_i, y_i) \vee 1$ to each sampled point. This amounts to a larger, weighted subsample from $\mathbb{P}_S$, and we can make the same correction to the estimates from the subsample. We see in Section 4.4 that for $c > 1$ the factor of two is replaced by a factor of $1 + \frac{1}{c}$.

In the case of large imbalance, most of the $\tilde{p}(x_i)$ are near 0 or 1. For $c > 1$, the marginal acceptance probability at x_i becomes

$$\mathbb{P}(z_i = 1 | x_i = x) = p(x)\big(c\big(1 - \tilde{p}(x)\big) \wedge 1\big) + \big(1 - p(x)\big)\big(c\tilde{p}(x) \wedge 1\big) \tag{19}$$

$$\approx (1 + c)p(x)\big(1 - p(x)\big), \tag{20}$$

where the approximation holds for $p(x) \approx \tilde{p}(x) \approx 0$ or 1. For $c = 1$, the marginal acceptance probability is $p(x)(1 - \tilde{p}(x)) + (1 - p(x))\tilde{p}(x) \approx 2p(x)(1 - p(x))$, so for $c > 1$ we take roughly $\frac{1+c}{2}$ times as many data points as for $c = 1$. For example, if $c = 5$, the subsample accepted is roughly 3 times as large, and the relative efficiency improves from $1/2$ to $5/6$.

Alternatively, if n is extremely large, even a small fraction of the full data set may still be more than we want. In that case, we can proceed as above with $c < 1$, or simply sample any desired number n_s of data points uniformly from the local case-control subsample.

4. Asymptotics of the local case-control estimate. We now turn to examining the asymptotic behavior of the local case-control estimate. We first establish consistency, assuming a consistent pilot estimate $\tilde{\theta}$. We expressly do not assume that the pilot estimate is independent of the data, since in some cases we may recycle into the subsample some of the data we used to calculate the pilot.

By assuming independence of $\tilde{\theta}$ and the data, we can obtain finer results about the asymptotic distribution of $\hat{\theta}$. We show it is asymptotically unbiased when $\tilde{\theta}$ is, and derive the asymptotic variance of the estimate. When the logistic regression model is correctly specified, the local case-control estimate has exactly twice the asymptotic variance of the MLE for the full data set.

4.1. *Preliminaries.* For better clarity of notation in this section, we will use the letter λ in place of $\tilde{\theta}$ to denote pilot estimates. Additionally, we drop the notation $\binom{1}{x}$ and absorb the constant term into x, so that $f_\theta(x) = \theta' x$. To avoid trivialities,

assume without loss of generality that there is no $v \in \mathbb{R}^p$ for which $\mathbb{E}|v'X| = 0$ (if not, we can discard redundant features).

For $\pi \in [0, 1]$ define the "soft hinge" function

$$h(\eta; \pi) = -\pi\eta + \log\bigl(1 + e^{\eta}\bigr), \tag{21}$$

and note that

$$\mathbb{E}\bigl[\rho(\theta; X, Y)|X = x\bigr] = h\bigl(\theta' x; p(x)\bigr). \tag{22}$$

As a function of η, h is positive and strictly convex, its magnitude is bounded by $1 + |\eta|$, and it has Lipschitz constant $\max(\pi, 1 - \pi) \le 1$. If $\pi < 1$, h diverges as $\eta \to \infty$, and if $\pi > 0$ h diverges as $\eta \to -\infty$.

As a function of λ, $a_\lambda(x, y) = |y - \frac{e^{\lambda' x}}{1+e^{\lambda' x}}| \in (0, 1)$ has Lipschitz constant $\le \|x\|$. Hence, $\bar{a}(\lambda) = \mathbb{E}a_\lambda(X, Y) \in (0, 1)$ with Lipschitz constant $\le \mathbb{E}\|X\|$. The marginal acceptance probability given x is

$$\hat{a}_\lambda(x) = \tilde{p}(x)\bigl(1 - p(x)\bigr) + \bigl(1 - \tilde{p}(x)\bigr)p(x) \in (0, 1). \tag{23}$$

Given pilot λ, the local case-control subsampling scheme effectively samples from the probability measure $\mathbb{P}_\lambda$, where

$$d\mathbb{P}_\lambda(x, y) = \frac{a_\lambda(x, y)\, d\mathbb{P}(x, y)}{\bar{a}(\lambda)}, \tag{24}$$

and $\bar{a}(\lambda) = \int a_\lambda(x, y)\, d\mathbb{P}(x, y)$ is the marginal probability of acceptance. Under this measure,

$$\operatorname{logit} \mathbb{P}_\lambda(Y = 1|X = x) = f(x) - \lambda' x. \tag{25}$$

Because $a_\lambda(x, y) \le 1$, if $g(X, Y)$ is integrable under $\mathbb{P}$ it is also integrable under any $\mathbb{P}_\lambda$.

Conditioning on X, we can write the population risk of the logistic regression parameters θ with respect to sampling measure $\mathbb{P}_\lambda$ as

$$R_\lambda(\theta) = \frac{-1}{\bar{a}(\lambda)} \int h\left(\theta' x; \frac{e^{f(x)-\lambda' x}}{1 + e^{f(x)-\lambda' x}}\right)\hat{a}_\lambda(x)\, d\mathbb{P}(x). \tag{26}$$

By Cauchy–Schwarz, the integrand in (26) is bounded by $2(1 + \|\theta\|\|x\|)$. If $\mathbb{E}\|X\| < \infty$, then, we may appeal to dominated convergence and take limits with respect to θ and λ inside the integral.

$R_\lambda(\theta)$ is strictly convex because the integrand is, and always has a unique population minimizer if there is no separating hyperplane in the population.

LEMMA 1. *Assume there is no v for which*

$$\mathbb{P}\bigl(Y = 0, v'X > 0\bigr) = \mathbb{P}\bigl(Y = 1, v'X < 0\bigr) = 0. \tag{27}$$

Henceforth, we refer to this assumption as nonseparability. *Then $R_\lambda(\theta)$ attains a unique minimum for every $\lambda \in \mathbb{R}^p$.*

Denote by $\widehat{R}_\lambda^{(0)}(\theta)$ the empirical risk on a local case-control subsample taken using the pilot estimate λ. Then

$$\widehat{R}_\lambda^{(0)}(\theta) = -\left(\sum_{i=1}^n z_i\right)^{-1} \sum_{i=1}^n z_i\big[y_i\theta' x_i - \log\big(1+e^{\theta' x_i}\big)\big]. \tag{28}$$

It will be somewhat simpler to replace the random subsample size $\sum_{i=1}^n z_i$ with its expectation $n\bar{a}(\lambda)$. Define

$$\widehat{R}_\lambda(\theta) = -\frac{1}{n\bar{a}(\lambda)} \sum_{i=1}^n z_i\big[y_i\theta' x_i - \log\big(1+e^{\theta' x_i}\big)\big]. \tag{29}$$

Since minimizing (28) with respect to θ is equivalent to minimizing (29), the two are equivalent for our purposes.

If the unadjusted parameters $\hat{\theta}_S$ minimize $\widehat{R}_\lambda$, the local case-control estimate $\hat{\theta} = \hat{\theta}_S + \lambda$ is an M-estimator minimizing $\widehat{Q}_\lambda(\theta) = \widehat{R}_\lambda(\theta - \lambda)$. We use analogous notation for the population version:

$$Q_\lambda(\theta) = R_\lambda(\theta - \lambda). \tag{30}$$

For any given pilot estimate λ and large n, we expect

$$\hat{\theta} \approx \arg\min_\theta Q_\lambda(\theta). \tag{31}$$

Define the right-hand side of (31) to be $\bar{\theta}(\lambda)$, the large-sample limit of local case-control sampling with pilot estimate fixed at λ. The best linear predictor for the original population corresponds to the case $\lambda = 0$ (uniform subsampling), that is, $\theta^* = \bar{\theta}(0)$. Consistency means that for large n, $\hat{\theta} \xrightarrow{p} \theta^*$.

Recall that if the model is correctly specified with true parameters θ_0, then $\bar{\theta}(\lambda) = \theta_0$ for *any* fixed pilot estimate λ. Minimizing $\widehat{Q}_\lambda$ therefore yields a consistent estimate. Unfortunately, in the misspecified case $\bar{\theta}(\lambda) \neq \bar{\theta}(0) = \theta^*$. In this sense, local case-control sampling with the pilot λ held fixed is in general *not* consistent for θ^*. However, we see below that it is consistent if $\lambda = \theta^*$.

PROPOSITION 2. *Assume* $\mathbb{E}\|X\| < \infty$, *that the classes are nonseparable*, *and that* $\theta^* = \bar{\theta}(0)$ *is the best linear predictor for the original measure* $\mathbb{P}$. *Then*

$$\theta^* = \arg\min_\theta Q_{\theta^*}(\theta) = \bar{\theta}\big(\theta^*\big). \tag{32}$$

In other words, if we could only choose our pilot perfectly, then the local case-control estimate would converge to θ^* as $n \to \infty$.

PROOF OF PROPOSITION 2. Write $p^*(x) = \frac{e^{\theta^{*\prime} x}}{1+e^{\theta^{*\prime} x}}$. The population optimality criterion for LCC with pilot λ is

$$0 = -\bar{a}(\lambda)\nabla_\theta R_\lambda(\theta - \lambda) \tag{33}$$

$$= -\mathbb{E}\big[X\rho'\big((\theta-\lambda)'X; X, Y\big)a_\lambda(X, Y)\big]. \tag{34}$$

Noting that $-\rho'(0; x, y) = y - \frac{1}{2}$, if we evaluate the above at $\lambda = \theta = \theta^*$, we obtain

$$\mathbb{E}\big[X\big(Y - \tfrac{1}{2}\big)a_\lambda(X, Y)\big] = \tfrac{1}{2}\mathbb{E}\big[X\big(p(X)\big(1 - p^*(X)\big) - \big(1 - p(X)\big)p^*(X)\big)\big] \tag{35}$$

$$= \tfrac{1}{2}\mathbb{E}\big[X\big(Y - p^*(X)\big)\big] \tag{36}$$

which is exactly half the population score (11) for the original population. Since θ^* optimizes the risk for the original population, this value is 0. □

There is an intuitive explanation of this result: in $\mathbb{P}_{\theta^*}$, the *acceptance probabilities* are $p^*(X)$ if $Y = 0$ and $1 - p^*(X)$ if $Y = 1$; hence they play the same role as the *pseudoresiduals* $Y - p^*(X)$ did in the original measure $\mathbb{P}$. For example, the point $(x, 0)$ would contribute $p^*(x)x$ to the gradient if we evaluated the full-sample score at θ^*. Evaluating the subsample score at 0, the same point now contributes $\frac{1}{2}x$ to the score—but only if it is accepted, which occurs with probability exactly $p^*(x)$. So, in essence, the subsampling stands in for the reweighting that we otherwise would have done when fitting our logistic regression to the full sample.

Of course, in practice we never have a perfect pilot—if we did we would not need to estimate θ^*—but Proposition 2 suggests that if λ is near θ^*, minimizing $\widehat{Q}_\lambda$ yields a good estimate. In fact, we will see that if $\lambda \xrightarrow{p} \theta^*$ then $\hat{\theta} \xrightarrow{p} \theta^*$ as well.

4.2. *Consistency.* For our asymptotic results, assume we have an infinite reservoir $(x_1, y_1), (x_2, y_2), \ldots$ of i.i.d. pairs, a sequence of i.i.d. $U(0, 1)$ variables $u_1, u_2, \ldots$ for making accept–reject decisions, and a sequence of pilot estimates $\lambda_1, \lambda_2, \ldots$. The λ_n are possibly dependent upon the data, but the u_i are assumed to be independent of everything else.

$\hat{\theta}_n$ is the local case-control estimate, computed using pilot λ_n, data $\{(x_i, y_i)\}_{i=1}^n$, and accept–reject decisions $z_i = \mathbf{1}_{u_i \le a_{\lambda_n}(x_i, y_i)}$.

The main result of this section is that if the pilot estimate λ_n is consistent for θ^*, then so is $\hat{\theta}_n$. The details are somewhat technical, especially the proof of Proposition 3, but the main idea is that if $\lambda_n \xrightarrow{p} \theta^*$, then for large n

$$\widehat{Q}_{\lambda_n} \approx Q_{\theta^*} \tag{37}$$

in the appropriate sense. $\widehat{Q}_{\lambda_n}$ is what the local case-control estimate actually minimizes, whereas the last function is minimized by θ^*, our ultimate target.

First, we establish pointwise convergence.

PROPOSITION 3. *If $\mathbb{E}\|X\| < \infty$ and $\lambda_n \xrightarrow{p} \lambda_\infty$, then for each $\theta \in \mathbb{R}^p$,*

$$\widehat{Q}_{\lambda_n}(\theta) \xrightarrow{p} Q_{\lambda_\infty}(\theta). \tag{38}$$

Because we avoid assuming independence between the pilot λ_n and the data (x_i, y_i), the proof is technical and is deferred to the Appendix. The proof relies on the coupling of the acceptance decisions z_i for different pilot estimates through u_i. With this coupling, two nearby pilot estimates will differ on very few accept–reject decisions.

Because neither $\widehat{Q}_{\lambda_n}(\theta)$ nor $Q_{\lambda_\infty}(\theta)$ changes very fast, pointwise convergence also implies uniform convergence on compacts.

PROPOSITION 4. *If* $\mathbb{E}\|X\| < \infty$ *and* $\lambda_n \xrightarrow{p} \lambda_\infty$, *then for compact* $\Theta \subseteq \mathbb{R}^p$,

$$\sup_{\theta\in\Theta} \left| \widehat{Q}_{\lambda_n}(\theta) - Q_{\lambda_\infty}(\theta) \right| \xrightarrow{p} 0. \tag{39}$$

PROOF. Define

$$F_n(\theta) = \widehat{Q}_{\lambda_n}(\theta) - Q_{\lambda_\infty}(\theta). \tag{40}$$

By Proposition 3, $F_n(\theta) \xrightarrow{p} 0$ pointwise. Next, we show it is Lipschitz. The integrand in (35) is x times two factors each bounded by ± 1, hence

$$\left\| \bar{a}(\lambda_\infty) \nabla_\theta Q_{\lambda_\infty} \right\| \le \int \|x\| \, d\mathbb{P}(x) = \mathbb{E}\|X\|. \tag{41}$$

Similarly for $\widehat{Q}_{\lambda_n}$, we have

$$\nabla_\theta \widehat{Q}_{\lambda_n} = -\frac{1}{n\bar{a}(\lambda_n)} \sum_{i=1}^{n} z_i \left(y_i - \frac{e^{(\theta-\lambda_n)' x_i}}{1 + e^{(\theta-\lambda_n)' x_i}} \right) x_i \tag{42}$$

so that

$$\sup_\theta \|\nabla_\theta \widehat{Q}_{\lambda_n}\| \le \bar{a}(\lambda_n)^{-1} \frac{1}{n} \sum_{i=1}^{n} \|x_i\| \xrightarrow{p} \bar{a}(\lambda_\infty)^{-1} \mathbb{E}\|X\|. \tag{43}$$

It follows that, with probability tending to 1, $F_n(\theta)$ has Lipschitz constant less than $c = 3\bar{a}(\lambda_\infty)^{-1}\mathbb{E}\|X\|$.

Now, for any $\varepsilon > 0$, we can cover Θ with finitely many Euclidean balls of radius $\delta = \varepsilon/c$, centered at $\theta_1, \ldots, \theta_{M(\varepsilon)}$. Let $A_n(\varepsilon)$ be the event that F_n has Lipschitz constant less than c and

$$\sup_{1\le j\le M(\varepsilon)} \left| F_n(\theta_j) \right| < \varepsilon. \tag{44}$$

On $A_n(\varepsilon)$, we have $\sup_{\theta\in\Theta} |F_n(\theta)| < 2\varepsilon$, and $\mathbb{P}(A_n(\varepsilon)) \to 1$ as $n \to \infty$. □

Finally, we come to the main result of the section, in which we prove that the local case-control estimate is consistent when the pilot is. Because the functions are strictly convex, we can ignore everything but a neighborhood of θ^*.

THEOREM 5. *Assume* $\mathbb{E}\|X\| < 0$ *and the classes are nonseparable. If* $\lambda_n \xrightarrow{p} \theta^*$ *then the local case-control estimate* $\hat{\theta}_n \xrightarrow{p} \theta^*$ *as well.*

PROOF. Let $\Theta \subseteq \mathbb{R}^p$ be any compact set with θ^* in its interior, and let

$$\varepsilon = \inf_{\theta \in \partial\Theta} Q_{\theta^*}(\theta) - Q_{\theta^*}(\theta^*) > 0, \tag{45}$$

where the strict inequality follows from strict convexity. Uniform convergence implies that with probability tending to 1,

$$\sup_{\theta \in \Theta} \left| \widehat{Q}_{\lambda_n}(\theta) - Q_{\theta^*}(\theta) \right| < \varepsilon/2 \tag{46}$$

which implies in turn that

$$\inf_{\theta \in \partial\Theta} \widehat{Q}_{\lambda_n}(\theta) > \widehat{Q}_{\lambda_n}(\theta^*). \tag{47}$$

Whenever this is the case, the strictly convex function $\widehat{Q}_{\lambda_n}$ has a unique minimizer in the interior of Θ. Since Θ was arbitrary, we can take its diameter to be less than any $\delta > 0$. Hence, $\hat{\theta}_n \xrightarrow{p} \theta^*$. $\square$

4.3. *Asymptotic distribution.* In this section, we derive the asymptotic distribution of the local case-control logistic regression estimate, in the same asymptotic regime as the previous section. To prove our results here, we assume the pilot estimate λ_n is independent of our data set. This would not be the case if our pilot were based on a subsample of the data (the procedure we use for all our simulations), but it could hold if the pilot came from a model fitted to data from an earlier time period.

The main result of this section is that if the logistic regression model is correctly specified and the pilot is consistent, the asymptotic covariance matrix of the local case-control estimate for θ is exactly twice the asymptotic covariance matrix of a logistic regression performed on the entire data set. For the results in this section, we will need $\mathbb{E}\|X\|^2 < \infty$.

It will be convenient to give names to some recurring quantities. First, we have seen that if $\mathbb{E}\|X\| < \infty$ we can differentiate $Q_\lambda(\theta)$ inside the integral to obtain the gradient of the population risk:

$$G(\theta, \lambda) \triangleq -\bar{a}(\lambda) \nabla_\theta Q_\lambda(\theta) \tag{48}$$

$$= \int \left(\frac{e^{f(x) - \lambda' x}}{1 + e^{f(x) - \lambda' x}} - \frac{e^{(\theta - \lambda)' x}}{1 + e^{(\theta - \lambda)' x}} \right) \hat{a}_\lambda(x) x \, d\mathbb{P}(x). \tag{49}$$

Whereas G is the expectation of the logistic regression score with respect to $\mathbb{P}_\lambda$, we can also define its covariance matrix:

$$J(\theta, \lambda) \triangleq \mathrm{Var}_\lambda \left[\left(Y - \frac{e^{(\theta - \lambda)' X}}{1 + e^{(\theta - \lambda)' X}} \right) X \right]. \tag{50}$$

When $\mathbb{E}\|X\|^2 < \infty$, $J(\theta, \lambda) < \infty$, and is continuous in θ and λ by dominated convergence.

Since the derivatives of the integrand in (48) are uniformly bounded by $2\|x\|^2$, dominated convergence implies we can again differentiate inside the integral. Differentiating with respect to θ we obtain

$$H(\theta,\lambda) \triangleq -\bar{a}(\lambda)\nabla_\theta^2 Q_\lambda(\theta) \tag{51}$$

$$= \int \frac{e^{(\theta-\lambda)'x}}{(1+e^{(\theta-\lambda)'x})^2}\left(\frac{e^{\lambda'x}+e^{f(x)}}{(1+e^{\lambda'x})(1+e^{f(x)})}\right)xx'\,d\mathbb{P}(x). \tag{52}$$

Here, the integrand is dominated by xx', so dominated convergence again applies and thus we see that H is continuous in θ and λ. $H(\theta,\lambda) \succ 0$ for any θ,λ since we have assumed there is no nonzero v for which $\mathbb{E}|v'X| = 0$. Finally, define the matrix of crossed partials:

$$C(\theta,\lambda) \triangleq \nabla_\lambda G(\theta,\lambda). \tag{53}$$

To be concrete, $C_{i,j} = \frac{\partial^2}{\partial\theta_i\partial\lambda_j}Q_\lambda(\theta)$. Continuity of C again follows from noting the derivative of the integrand in (48) with respect to λ is dominated by $8\|x\|^2$.

To begin, we consider the behavior of $\bar{\theta}(\lambda)$ for λ near θ^*. By Proposition 2, we have $G(\theta^*,\theta^*) = 0$. Since $H(\theta,\lambda) \succ 0$, we can apply the implicit function theorem to the relation $G(\bar{\theta}(\lambda),\lambda) = 0$ to obtain

$$\bar{\theta}(\lambda) = \theta^* + H\big(\theta^*,\theta^*\big)^{-1}C\big(\theta^*,\theta^*\big)\big(\lambda-\theta^*\big) + o\big(\|\lambda-\theta^*\|\big). \tag{54}$$

By standard M-estimator theory, if we fix λ and send $n\to\infty$ the coefficients of a logistic regression performed on a sample of size $|S|$ from $\mathbb{P}_\lambda$ would be asymptotically normal with covariance matrix

$$\frac{1}{|S|}H\big(\bar{\theta}(\lambda),\lambda\big)^{-1}J\big(\bar{\theta}(\lambda),\lambda\big)H\big(\bar{\theta}(\lambda),\lambda\big)^{-1}. \tag{55}$$

In light of this and the fact that $|S| \approx \bar{a}(\lambda)n$, we might predict the following.

THEOREM 6. *Assume $\mathbb{E}\|X\|^2 < \infty$. If $\lambda_n \xrightarrow{p} \theta^*$ independently of the data, then*

$$\sqrt{n}\big(\hat{\theta}_n - \bar{\theta}(\lambda_n)\big) \xrightarrow{\mathcal{D}} N\big(0,\bar{a}\big(\theta^*\big)^{-1}\Sigma\big) \tag{56}$$

with $\Sigma = H(\theta^*,\theta^*)^{-1}J(\theta^*,\theta^*)H(\theta^*,\theta^*)^{-1}$.

Again, we defer the proof to the Appendix. We can combine (56) with (54) to immediately obtain the following reassuring facts.

COROLLARY 7. *Assume $\mathbb{E}\|X\|^2 < \infty$ and λ_n is a sequence of pilot estimators given independently of the data. Then*:

(a) *If λ_n is $\sqrt{n}$-consistent, so is $\hat{\theta}_n$.*
(b) *If λ_n is asymptotically unbiased, so is $\hat{\theta}_n$.*

(c) *If* $\sqrt{n}(\lambda_n - \theta^*) \xrightarrow{\mathcal{D}} N(0, V)$ *then* $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\mathcal{D}} N(0, \Sigma)$ *with*

$$\Sigma = H^{-1}\big(CVC' + \bar{a}^{-1}J\big)H^{-1}. \tag{57}$$

In (57), *we have suppressed the arguments of* θ^* *in* $H, C, \bar{a}$ *and* J.

The first term in (57) characterizes the contribution of conditional bias (given λ_n) to the overall variance, and the second is the contribution of conditional variance.

In the special case where logistic regression model is correctly specified, we have the following.

THEOREM 8. *Assume the logistic regression model is correct and let* $\frac{1}{n}\Sigma_{\text{full}}$ *be the asymptotic variance of the MLE for the full sample. Then if* $\mathbb{E}\|X\|^2 < \infty$ *and* $\lambda_n \xrightarrow{p} \theta_0$ *independently of the data, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N\big(0, a(\theta_0)^{-1}\Sigma\big) = N(0, 2\Sigma_{\text{full}}). \tag{58}$$

Hence, although the size of a local case-control subsample is roughly $n\bar{a}(\lambda)$, the variance of $\hat{\theta}$ is the same as if we took a simple random sample of size $n/2$ from the full data set. In other words, each point sampled is worth about $\frac{1}{2\bar{a}(\lambda_n)}$ points sampled uniformly.

PROOF OF THEOREM 8. If logistic regression is correctly specified for $\mathbb{P}$, it is also for $\mathbb{P}_\lambda$, regardless of λ, so $\bar{\theta}(\lambda) \equiv \theta_0$. Furthermore, by standard maximum likelihood theory $J(\theta_0, \lambda) = H(\theta_0, \lambda)^{-1}$ for each λ. Therefore, (56) specializes to

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N\big(0, \bar{a}(\theta_0)^{-1}H(\theta_0, \theta_0)^{-1}\big). \tag{59}$$

But

$$H(\theta, \lambda) = \bar{a}(\lambda)^{-1}\int \left[\frac{e^{(\theta-\lambda)'x}}{(1+e^{(\theta-\lambda)'x})^2}\right]\left[\frac{e^{\lambda'x}+e^{f(x)}}{(1+e^{\lambda'x})(1+e^{f(x)})}\right]xx'\,d\mathbb{P}(x). \tag{60}$$

If $f(x) = \theta_0'x$ and $\lambda = \theta_0$, then (60) simplifies to

$$H(\theta_0, \theta_0) = \bar{a}(\theta_0)^{-1}\frac{1}{2}\int \frac{e^{\theta_0'x}}{(1+e^{\theta_0'x})^2}xx'\,d\mathbb{P}(x) \tag{61}$$

$$= \bar{a}(\theta_0)^{-1}\frac{1}{2}H(\theta_0, 0) \tag{62}$$

$$= \big(2\bar{a}(\theta_0)\Sigma_{\text{full}}\big)^{-1}. \tag{63}$$

□

This result is surprisingly simple. No characterization like Theorem 8 is available for the case-control and weighted case-control estimates, whose variances are not simple scalar multiples of Σ_{full}.

We can offer a simple heuristic argument for Theorem 8, similar to that of Proposition 2. In $\mathbb{P}_{\theta_0}$, the acceptance probability $\hat{a}_\lambda(x)$ for an observation at x is $2p(x)(1-p(x))$, and given that it is accepted it contributes $\frac{1}{4}xx'$ to the observed information. In the full sample, a point at x is always accepted but contributes less, $p(x)(1-p(x))xx'$, to the observed information. Again, the sampling probability stands in for the reweighting we would have done in the full sample. If $p(x)(1-p(x))$ is very small, we are discarding most of the data instead of keeping all of it and assigning it a tiny weight in the fit.

The practical meaning of Theorem 8 is that local case-control sampling is most advantageous when $\bar{a}(\theta_0) = \mathbb{E}(|Y - \tilde{p}(X)|)$ is small, that is, when Y is easy to predict throughout much of the covariate space. This can happen as a result of marginal or conditional imbalance, or both. Standard case-control sampling can also improve our efficiency in the presence of marginal imbalance, but unlike local case-control sampling, it does not exploit conditional imbalance. Hence, we would expect local case-control to outperform standard case-control most dramatically when the marginal imbalance is very high, as in the simulation of Section 5.2.

For data-dependent pilots, the efficiency picture is somewhat more complicated. For example, $\bar{\theta}(\lambda)$ is approximately a linear function of $\lambda - \theta^*$. Thus, if λ is unbiased but correlated with the noise in the data, we might get more or less variance relative to (58), depending on how this correlation interacts with C. If the model is correctly specified, it less clear whether an adversarially chosen pilot can affect the efficiency.

Either way, we do not anticipate serious problems from nonindependence. To stress-test our results against violations of independence, we expressly use a data-dependent pilot for all of our experiments: namely, a weighted case-control sample with sample points allowed to be recycled for the second-stage fit.

4.4. *Variance for a larger sample.* In Section 3.3, we proposed increasing the size of the local case-control subsample by multiplying all the acceptance probabilities $a(x, y)$ by a constant $c > 1$ and assigning weight $w = ca(x, y)$ when $ca(x, y) > 1$. We analyze the asymptotic variance here as a function of c. To simplify matters, suppose the model is correctly specified and λ is fixed at θ_0.

The weighted log-likelihood for the subsample and its derivatives are then

$$\ell_w(\theta) = \sum_{i=1}^{n} z_i w_i \big(y_i \theta' x_i - \log\big(1 + e^{\theta' x_i}\big)\big), \tag{64}$$

$$\nabla_\theta \ell_w(\theta) = \sum_{i=1}^{n} z_i w_i \big(y_i - p_\theta(x_i)\big) x_i, \tag{65}$$

$$\nabla^2_\theta \ell_w(\theta) = \sum_{i=1}^{n} z_i w_i p_\theta(x_i)\big(1 - p_\theta(x_i)\big) x_i x_i'. \tag{66}$$

Conditionally on x, there is a $p(x) \cdot (c(1-p(x)) \wedge 1)$ chance $y = z = 1$ and $w = c(1-p(x)) \vee 1$, where $p(x) = p_{\theta_0}(x)$. Similarly, there is a $(1-p(x)) \cdot (cp(x) \wedge 1)$ chance $y = 0, z = 1$, and $w = cp(x) \vee 1$. We immediately obtain

$$\begin{aligned} \mathbb{E}(yzw|x) &= cp(1-p), \\ \mathbb{E}(zw|x) &= 2cp(1-p), \\ \mathbb{E}\big(zw^2|x\big) &\le c(c+1)p(1-p). \end{aligned} \tag{67}$$

The expectation and variance of the score evaluated at 0 are

$$\mathbb{E}\nabla_\theta \ell_w(0) = n \int \mathbb{E}\big(zw(y-1/2)|x\big)x \, d\mathbb{P}(x) = 0, \tag{68}$$

$$J = \operatorname{Var}\big(\nabla_\theta \ell_w(0)\big) = n \int \mathbb{E}\big(z^2w^2(y-1/2)^2|x\big)xx' \, d\mathbb{P}(x) \tag{69}$$

$$= \frac{n}{4} \int \mathbb{E}\big(zw^2|x\big)xx' \, d\mathbb{P}(x) \preceq \frac{c(c+1)}{4}\Sigma_{\text{full}} \tag{70}$$

and the expected Hessian is

$$H = \mathbb{E}\nabla_\theta^2 \ell_w(0) = \frac{n}{4} \int \mathbb{E}(zw|x)xx' \, d\mathbb{P}(x) = \frac{c}{2}\Sigma_{\text{full}}^{-1}. \tag{71}$$

We have derived

$$H^{-1} J H^{-1} \preceq \left(1 + \frac{1}{c}\right)\Sigma_{\text{full}}. \tag{72}$$

For $c = 1$, we recover the factor of two from (58), but, for example, $c = 5$ we only pay 20% increased variance relative to the full sample.

5. Simulations. Here, we compare our method to standard weighted and unweighted case-control sampling for two-class Gaussian models like the one considered in Section 2.2. The standard case-control estimates use a 50–50 split between the two classes.

5.1. *Simulation* 1: *Two-class Gaussian, different variances*. We begin with a five-dimensional two-class Gaussian simulation where the classes have different covariance matrices. If $X|Y = y \sim N(\mu_y, \Sigma_y)$, then

$$\begin{aligned} \log \frac{\mathbb{P}(x|Y=1)}{\mathbb{P}(x|Y=0)} &= -\frac{1}{2}(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) \\ &\quad + \frac{1}{2}(x-\mu_0)'\Sigma_0^{-1}(x-\mu_0) + \text{const.} \end{aligned} \tag{73}$$

Equation (73) is linear if $\Sigma_1 = \Sigma_0$, and quadratic otherwise, so if the two covariance matrices were the same the linear logistic model would be correctly specified.

In this case the model is incorrectly specified, letting us compare the behavior of the different methods under model misspecification.

Take $\mathbb{P}(Y = 1) = 1\%$, $\mu_0 = 0$, and $\mu_1 = (1, 1, 1, 1, 4)'$. The covariance matrices are $\Sigma_0 = \text{diag}(1, 1, 1, 1, 9)$ and $\Sigma_1 = I_5$. Hence $f(x)$ is additive, but with a nonzero quadratic term in x_5.

For our simulation, we first generate a large ($n = 10^6$) sample from the population described above. Second, we obtain a pilot model using the weighted case-control method on $n_s = 1000$ data points. Next, we take a local case-control sample of size 1000 using that pilot model.

For comparison, we obtain standard case-control (CC) and weighted case-control (WCC) estimates. For the comparison estimators we do not use a sample of size 1000 again but rather use the total number of observations seen by the LCC model or the pilot model, roughly 2000, so the LCC estimate must pay for its pilot sample. We repeat this entire procedure 1000 times.

Table 2 shows the squared bias and variance of $\hat{\beta}$ over the 1000 realizations for each of the three methods. As expected, we face a bias-variance tradeoff in choosing between the WCC and CC methods, whereas the LCC method improves substantially on the bias of CC and the variance of WCC. Standard errors for both bias and variance are computed via bootstrapping the 1000 realizations.

More surprising is the fact that LCC enjoys smaller bias than WCC and smaller variance than CC, dominating the other two methods on both measures. The improvement in variance over the CC estimate is likely due to the conditional imbalance present in the sample, while the improvement in bias over the WCC estimate may come from the fact that the methods are only unbiased asymptotically and the LCC estimate is closer to its asymptotic limiting behavior.

TABLE 2

Estimated bias and variance of $\hat{\beta}$ for each sampling method. For $\hat{\beta} \in \mathbb{R}^p$, we define $\text{Bias}^2 = \|\mathbb{E}\hat{\beta} - \beta\|^2$ *and* $\text{Var} = \sum_{j=1}^{p} \text{Var}(\hat{\beta}_j)$

Simulation 1 ($\Sigma_0 \neq \Sigma_1 \Rightarrow$ *model misspecified*)

	$\widehat{\text{Bias}}^2$	(s.e.)	$\widehat{\text{Var}}$	(s.e.)
LCC	0.0049	(0.00031)	0.025	(0.00059)
WCC	0.023	(0.0022)	0.16	(0.0038)
CC	0.15	(0.0016)	0.043	(0.00096)

Simulation 2 ($\Sigma_0 = \Sigma_1 \Rightarrow$ *model correct*)

	$\widehat{\text{Bias}}^2$	(s.e.)	$\widehat{\text{Var}}$	(s.e.)
LCC	0.0037	(0.0083)	0.039	(0.00045)
WCC	0.59	(0.064)	1.7	(0.017)
CC	0.06	(0.042)	0.87	(0.0086)

5.2. *Simulation* 2: *Two-class Gaussian, same variance.* Next, we simulate a two-class Gaussian model with each class having the same variance, so that the true log-odds function f is linear. We also increase the dimension to 50 for this simulation.

Since the model is now correctly specified, all three methods are asymptotically unbiased. However, in this case we introduce more substantial conditional imbalance, to demonstrate the variance-reduction advantages of local case-control sampling in that setting.

For this example, $\mathbb{P}(Y = 1) = 10\%$, $\mu_1 = \binom{1_{25}}{0_{25}}$, $\mu_0 = 0_{50}$, and $\Sigma_0 = \Sigma_1 = I_{50}$. We repeat the procedure from Section 5.1, now with $n_s = 10^4$. Instead of generating a full sample, the full data set is implicit and we sample directly from $\mathbb{P}_S$.

In this example, the difference between the methods is more dramatic. Table 2 shows the squared bias and variance of the three methods. Here, local case-control enjoys substantially better bias than the other two methods, improving on CC more than twenty-fold. For the correct pilot model, $\bar{a}(\theta_0)$ is roughly 0.005, so the local case-control subsample size is around $n/200$. Since the model is correctly specified, the variance is roughly twice that of logistic regression on the full sample of size n. In other words, local case-control subsampling is roughly 100 times more efficient than uniform subsampling.

Asymptotically, all three methods are unbiased but it appears that LCC again enjoys a smaller bias in finite samples.

6. Web spam data set. Relative to standard case-control sampling, local case-control sampling is especially well-suited for data sets with significant conditional imbalance, that is, data sets in which y_i is easy to predict for most x_i.

One such application is spam filtering. To demonstrate the advantages of local case-control sampling and compare asymptotic predictions to actual performance, we test our method on the Web Spam data available on the LIBSVM website[3] and originally from Webb, Caverlee and Pu (2006). The data set contains 350,000 web pages, of which about 60% are labeled as "web spam," that is, web pages designed to manipulate search engines rather than display legitimate content. This data set is marginally balanced, though as we will see the conditional imbalance is considerable.

As features, we use frequency of the 99 unigrams that appeared in at least 200 documents, log-transformed with an offset so as to reduce skew in the features. In this data set, the downsampling ratio $\bar{a}$ is around 10%, that is, when using a good pilot we will retain about 10% of the observations.

Since we only have a single data set, we use subsampling as a method to assess the sampling distribution of our estimators. In each of 100 replications, we begin by taking a uniform subsample of size $n = 100{,}000$ from the population of 350,000

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

documents. After obtaining 100 data sets of size $n = 100{,}000$, we use the same procedure as we used in our two simulations with $n_S = 10{,}000$.

Our asymptotic theory predicts that the variance of the local case-control sampling estimate of θ should be a little more than twice the variance using the full sample (more because the model is misspecified and our pilot has some variance). Because the full sample is close to marginally balanced, the standard case-control sampling methods should do about as well as a uniform subsample of size 20,000—that is, they should have variance roughly 5 times that of the full sample.

Note that 20,000 is roughly twice the size of the local case-control sample, since we are counting the pilot sample against the local case-control method. If we had a readily available pilot model, as we would in many applications, it would be more relevant to give the CC and WCC methods access to only 10,000 data points, doubling their variance relative to the observed variance in this experiment.

The theoretical predictions come reasonably close in this experiment, as shown in Figure 3. The horizontal axis indexes each of the 100 coefficients to be fit (there are 99 covariates and an intercept), and the vertical axis gives the variance of each estimated coefficient, relative to the variance of the same coefficient in a model fitted to the full sample.

The magnitude of our improvement over standard case-control sampling is substantial here, but could be much larger in a data set with an even stronger signal. The key point is that standard case-control methods have no way to exploit conditional imbalance, so the more there is, the more local case-control dominates the other methods.

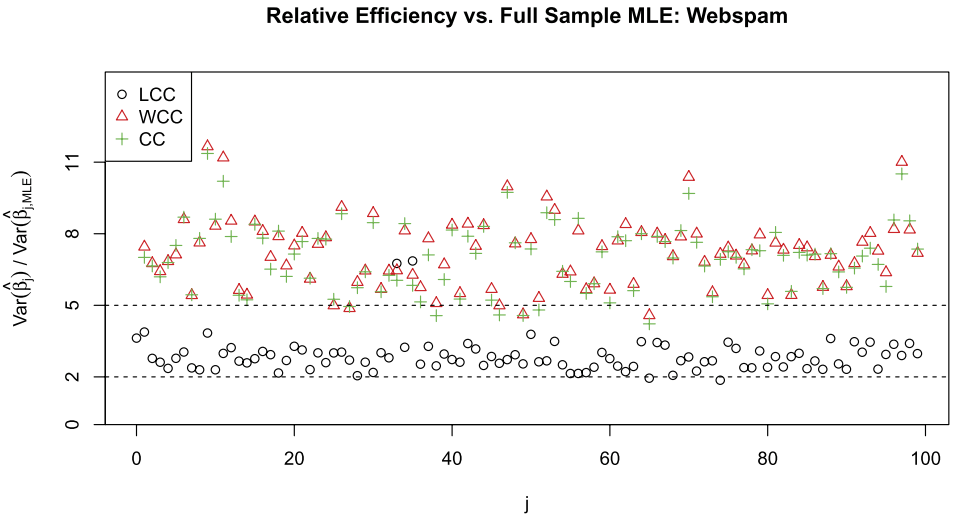

FIG. 3. *Relative variance of coefficients for different subsampling methods. The theoretical predictions* (2× *variance for local case-control*, 5× *variance for standard*) *are reasonably close to the mark, though a bit optimistic.*

7. Discussion. We have shown that in imbalanced logistic regression, we can speed up computation by subsampling the data in a biased fashion and making a post-hoc correction to the coefficients estimated in the subsample. Standard case-control sampling is one such scheme, but it has two main flaws: it has no way to exploit conditional imbalance, and when the model is misspecified it is inconsistent for the population risk minimizer.

Local case-control sampling generalizes standard case-control sampling to address both flaws, subsampling with a bias that is allowed to depend on both x and y. When the pilot is consistent, our estimate is consistent even under misspecification, and if the model is correct then local case-control sampling has exactly twice the asymptotic variance of logistic regression on the full data set. Our simulations suggest that local case-control performs favorably in practice.

7.1. *Translating computational gains to statistical gains.* In the Introduction, we motivated our inquiry by identifying four ways that computational gains can translate to statistical ones. Specifically, we suggested that computational savings can:

(1) enable us to experiment with and prototype a variety of models, instead of trying only one or two,
(2) allow us to refit our models more often to adapt to changing conditions,
(3) allow for cross-validation, bagging, boosting, bootstrapping, or other computationally intensive statistical procedures or
(4) open the door to using more sophisticated statistical techniques on a compressed data set.

It is relatively clear how our proposed method can help with points (1) and (2). As for point (3), faster fitting procedures can directly speed up straightforward resampling techniques like bootstrapping or cross-validation, possibly making them feasible at scales where they previously were not. We discuss in Section 7.2 how it can also help with boosting.

The basic method as we have described it above does not deliver on point (4), because the pilot model and second-stage model are the same. However, an extension of our method can help, which we discuss below.

There is no reason in principle why the pilot model must be linear, or belong to the same model class as the model we fit to the local case-control sample. We can use any pilot predictions $\tilde{p}(x) = \frac{e^{\tilde{f}(x)}}{1+e^{\tilde{f}(x)}}$ in the sampling algorithm, and then model the log-odds in the subsample quite flexibly—by a GAM, kernel logistic regression, random forests or any other method—so long as we can use offsets $-\tilde{f}(x_i)$ in the second-stage procedure. For example, we could use as our pilot fit a simple model with a few important variables explaining most of the response, and in the second stage estimate more complex models refining the first.

Formally, our theoretical results may not cover this use. Suppose the second-stage model can be written as a logistic regression after some basis expansion.

Then consistency of the second-stage estimate requires either that the pilot be consistent (the new variables contribute nothing to the population fit) or that the second-stage model be correctly specified. If neither of these assumptions holds approximately, then our estimate could be biased—though perhaps not as biased as case-control sampling, which is a special case of local case-control with an intercept-only pilot.

If we are prototyping, guarantees of consistency may not be a high priority. If they are, then as with case-control sampling, we can repair the inconsistency of the local case-control estimate by using a Horvitz–Thompson estimator with weights $a_{\tilde{\theta}}(x_i, y_i)^{-1}$. This may come at a cost of some added variance. It would be interesting to examine the bias of local case-control and the variance of weighted local case-control in this more general problem setting.

7.2. *Extensions*. This work suggests extensions in several directions, described below.

Indifference point other than 50%. In some applications (e.g., diagnostic medical screening), a false negative may be more costly than a false positive, or vice-versa. One of the implications of the discussion in Section 2.2 is that the Bernoulli log-likelihood implicitly places most emphasis on approximating the log-odds well near the 0 (50% probability) level curve, which may not be appropriate if the decision boundary relevant to our application is at 10%. In general, we would expect to obtain a better model in the large-n limit if we target the decision boundary we care most about.

In a sense, the reason that standard case-control sampling performed so badly in Example 2 of Section 2.2 is that it targeted a level curve of $\mathbb{P}(Y = 1|X = x)$ other than 50%. Specifically, it targeted the level curve corresponding to 50% in the subsampling population for equal-sampled case-control sampling, which corresponds to the marginal $\mathbb{P}(Y = 1)$ level curve in the original population.

What happened by accident in Example 2 need not always be one, and it would be interesting to generalize our procedure so as to target any chosen decision threshold. More generally still, our indifference point could depend on our features x—in online advertising, for instance, some advertisers may be willing to pay more per click than others.

Boosting. In Section 7.1 we suggested using offsets to obtain a complex second-stage fit. Alternatively, we can obtain any fitted log-odds function $f_s(x)$ for the sample and simply add it to the pilot $\tilde{f}(x)$ to obtain an estimate for $f(x)$.

This observation suggests the possibility of iteratively fitting a "base model" to the subsample, then adding it to $\tilde{f}(x)$ to obtain a new pilot for the next iteration. Indeed, that iterative algorithm is closely related to the AdaBoost algorithm of Freund and Schapire (1997). Even more similarly to AdaBoost, we could weight each point by $|y_i - \tilde{p}(x_i)|$ instead of sampling it with that probability.

Friedman, Hastie and Tibshirani (2000) show that the AdaBoost algorithm can be thought of as fitting a logistic regression model additive in base learners. In AdaBoost, the function $F_M(x) = \sum_{m=1}^{M} f_m(x)$ simply records the number of classifiers f_m classifying x as belonging to class $+1$ minus the number classifying it as class -1, and Friedman et al. show that $\frac{1}{2}F_M(x)$ can be thought of as approximating the log-odds of $Y = +1$ given $X = x$.

The difference is that while AdaBoost weights the point (x_i, y_i) by $e^{(2y_i-1)F_m(x_i)}$, the local case-control version would use weights

$$|y_i - p_M(x_i)| = \frac{e^{y_i F_m(x_i)}}{1 + e^{F_m(x_i)}} = \frac{e^{(2y_i-1)F_m(x_i)}}{1 + e^{(2y_i-1)F_m(x_i)}}. \tag{74}$$

Operationally, this alternative weighting scheme limits the influence of "outliers," that is, hard-to-classify points that can unduly drive the AdaBoost fit.

Logistic regression with regularization. In high-dimensional settings, lasso- or ridge-penalized logistic regressions are often preferable to standard logistic regression, the model considered here. One could use local case-control sampling with a regularized version of logistic regression, but our asymptotic results might need revisiting in such a case—especially in a high-dimensional asymptotic regime [$p \gg n$ or $p/n \to \gamma \in (0, \infty)$]. Since the high-dimensional setting is important in modern statistics and machine learning, this bears further investigation.

Other generalized linear models. One way of viewing the method is as a way of "tilting" the conditional distribution of Y by a linear function of X in the natural parameter space so as to enrich our subsample for more informative observations. We could use similar tricks on other GLMs.

For instance, suppose we are given a Poisson variable with natural parameter $\eta = \log \mathbb{E}Y$. By sampling with acceptance probability proportional to $e^{\xi Y}$, we obtain (conditional on acceptance) a Poisson with natural parameter $\eta + \xi$. Since Poisson variables with larger means carry more information, this could yield a substantial improvement over uniform subsampling.

If our data arise from a Poisson GLM with $\eta(x) \approx \alpha + \beta' x$, we could generalize the local case-control scheme by sampling (x_i, y_i) with probability proportional to $\exp\{(\xi_0 - \alpha - \beta' x_i)y_i\}$, where the extra parameter ξ_0 guarantees that we always tilt the conditional mean of y_i upward. Similar generalizations may apply for multinomial logit and survival models.

APPENDIX A: PROOF OF LEMMA 1 (UNIQUENESS OF θ^*)

Because $R_\lambda(\theta)$ is strictly convex, it is sufficient to show that $R_\lambda(\theta) \to \infty$ as $\theta \to \infty$ in any direction.

Assume w.l.o.g. there is some neighborhood $N \subseteq \mathbb{R}^p$ for which

$$\inf_{x \in N} \frac{\theta' x}{\|\theta\|} = \varepsilon > 0, \qquad \mathbb{P}(X \in N) > 0 \quad \text{and} \quad \mathbb{P}(Y = 1 | X \in N) = \pi_N < 1. \tag{75}$$

$h(\eta; \pi_N)$ is linear in its second argument, and is increasing for sufficiently large η. Thus, for large enough $\|\theta\|\varepsilon$, the population risk for $\mathbb{P}$ is

$$R(\theta) = \int h\big(\theta' x;\, p(x)\big)\, d\mathbb{P}(x) \tag{76}$$

$$\geq \int_N h\big(\|\theta\|\varepsilon;\, p(x)\big)\, d\mathbb{P}(x) \tag{77}$$

$$= h\big(\|\theta\|\varepsilon; \pi_N\big)\mathbb{P}(X \in N) \to \infty. \tag{78}$$

$\mathbb{P}_\lambda \gg \mathbb{P}$ for any λ, so (75) holds for $\mathbb{P}_\lambda$ with the same N (but a different $\pi_N < 1$). Thus, we can repeat the same argument with $\mathbb{P}$ replaced by $\mathbb{P}_\lambda$.

APPENDIX B: PROOF OF PROPOSITION 3 (POINTWISE CONVERGENCE)

Fix θ and begin by writing

$$\ell_i^\lambda = y_i(\theta - \lambda)' x_i - \log\big(1 + e^{(\theta-\lambda)' x_i}\big). \tag{79}$$

Let z_i^λ be the Bernoulli selection decisions, generated by comparing mutually independent $u_i \sim U(0,1)$ to the threshold $a_\lambda(x_i, y_i)$. The z_i^λ are independent conditional on λ and the data. Also, write $q_i^\lambda = z_i^\lambda \ell_i^\lambda$, so that $\widehat{Q}_\lambda(\theta) = \frac{-1}{n\bar{a}(\lambda)} \sum_{i=1}^n q_i^\lambda$.

By the Cauchy–Schwarz inequality, we have

$$\big|\ell_i^\lambda\big| \leq 1 + |\|\theta - \lambda\|\|x_i\|. \tag{80}$$

Now, for $\delta > 0$ define $\Lambda_\delta = \{\lambda : \|\lambda - \lambda_\infty\| < \delta\}$. For $\lambda \in \Lambda_1$, we have

$$\big|q_i^\lambda\big| \leq m_i \triangleq 1 + \big(\|\theta - \lambda_\infty\| + 1\big)\|x_i\| \tag{81}$$

which is integrable by assumption. Finally let $\mathbb{E}_n$ denote an average taken over indices $i = 1, \ldots, n$, that is, $\mathbb{E}_n f = \frac{1}{n}\sum_{i=1}^n f_i$. Then

$$\widehat{Q}_{\lambda_n}(\theta) - Q_{\lambda_\infty}(\theta) = \bar{a}(\lambda_n)^{-1}\mathbb{E}_n q^{\lambda_n} - \bar{a}(\lambda_\infty)^{-1}\mathbb{E} q^{\lambda_\infty}. \tag{82}$$

By continuity, $\bar{a}(\lambda_n) \xrightarrow{p} \bar{a}(\lambda_\infty) > 0$. Therefore, it suffices to show $\mathbb{E}_n q^{\lambda_n} \xrightarrow{p} \mathbb{E} q^{\lambda_\infty}$. Because $\mathbb{E}_n q^{\lambda_\infty} \xrightarrow{\text{a.s.}} \mathbb{E} q^{\lambda_\infty}$ by the law of large numbers, it suffices equally well to show that $\mathbb{E}_n q^{\lambda_n} - \mathbb{E}_n q^{\lambda_\infty} \xrightarrow{p} 0$.

Now fix $\varepsilon > 0$ and take K large enough that $\mathbb{E}(m\mathbf{1}_{m>K}) < \varepsilon$. For $\lambda_n \in \Lambda_1$ we have

$$\big|\mathbb{E}_n q^{\lambda_n} - \mathbb{E}_n q^{\lambda_\infty}\big| \leq \big|\mathbb{E}_n\big(q^{\lambda_n} - q^{\lambda_\infty}\big)\mathbf{1}_{m\leq K}\big| + 2\mathbb{E}_n m\mathbf{1}_{m>K}. \tag{83}$$

With probability one the second term is eventually less than 2ε. Further, for $\lambda_n \in \Lambda_\delta$, we have

$$|q_i^{\lambda_n} - q_i^{\lambda_\infty}| = \tfrac{1}{2}\big|\big(z_i^{\lambda_n} - z_i^{\lambda_\infty}\big)\big(\ell_i^{\lambda_n} + \ell_i^{\lambda_\infty}\big) + \big(z_i^{\lambda_n} + z_i^{\lambda_\infty}\big)\big(\ell_i^{\lambda_n} - \ell_i^{\lambda_\infty}\big)\big| \tag{84}$$

$$\leq \big|z_i^{\lambda_n} - z_i^{\lambda_\infty}\big| m_i + \delta\|x_i\|. \tag{85}$$

Now, write

$$d_i = \bigl|z_i^{\lambda_n} - z_i^{\lambda_\infty}\bigr| m_i \mathbf{1}_{m_i \le K}. \tag{86}$$

$z_i^{\lambda_n} \neq z_i^{\lambda_\infty}$ iff u_i lies between $a_{\lambda_n}(x_i, y_i)$ and $a_{\lambda_\infty}(x_i, y_i)$. Hence, conditionally on λ_n and the data, the d_i are mutually independent nonnegative random variables bounded by K with means

$$\mu_i = \bigl|a_{\lambda_n}(x_i, y_i) - a_{\lambda_\infty}(x_i, y_i)\bigr| m_i \mathbf{1}_{m_i \le K} < \delta K^2 \tag{87}$$

since $\nabla_\lambda a_\lambda(x_i, y_i) \le \|x_i\| < m_i$.

Continuing, we have

$$\bigl|\mathbb{E}_n\bigl(q^{\lambda_n} - q^{\lambda_\infty}\bigr)\mathbf{1}_{m \le K}\bigr| \le \mathbb{E}_n(d - \mu) + \mathbb{E}_n\mu + \delta\mathbb{E}_n\|x\|\mathbf{1}_{m \le K} \tag{88}$$

$$\le \mathbb{E}_n(d - \mu) + \delta K^2 + \delta K. \tag{89}$$

Conditioning on λ and $\{(x_i, y_i)\}$, the first term is a sum of independent zero-mean random variables that are bounded in absolute value by K. By Hoeffding's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} d_i - \mu_i\right| \ge \varepsilon \Bigm| \lambda_n, \{(x_i, y_i)\}\right) \le 2\exp[-n\varepsilon^2/(2K^2)]. \tag{90}$$

Since this bound is deterministic, the same applies to the unconditional probability that $\mathbb{E}_n(d - \mu)$ is large. Take $\delta = \varepsilon/(K + K^2)$. With probability tending to 1, $\lambda_n \in \Lambda_\delta$ and the event in (90) holds, in which case

$$\bigl|\mathbb{E}_n\bigl(q^{\lambda_n} - q^{\lambda_\infty}\bigr)\bigr| \le 4\varepsilon. \tag{91}$$

Since ε was arbitrary, the proof is complete.

APPENDIX C: PROOF OF THEOREM 6 [DISTRIBUTION OF $\hat{\theta} - \bar{\theta}(\lambda)$]

By the mean value theorem, we have for each n

$$\nabla_\theta \widehat{Q}_{\lambda_n}(\hat{\theta}_n) = \nabla_\theta \widehat{Q}_{\lambda_n}(\bar{\theta}(\lambda_n)) + \nabla_\theta^2 \widehat{Q}_{\lambda_n}(\phi_n)\bigl(\hat{\theta}_n - \bar{\theta}(\lambda_n)\bigr), \tag{92}$$

where ϕ_n is some convex combination of $\hat{\theta}_n$ and $\bar{\theta}(\lambda_n)$. Noting that the LHS is by definition 0 and rearranging, we obtain

$$\sqrt{n}\bigl(\hat{\theta}_n - \bar{\theta}(\lambda_n)\bigr) = \nabla_\theta^2 \widehat{Q}_{\lambda_n}(\phi_n)^{-1} \cdot \sqrt{n}\nabla_\theta \widehat{Q}_{\lambda_n}\bigl(\bar{\theta}(\lambda_n)\bigr). \tag{93}$$

If we can show the first factor tends in probability to $\nabla_\theta^2 Q_{\theta^*}(\theta^*)^{-1}$ and the second tends in distribution to $N(0, \bar{a}(\theta^*)^{-1} J(\theta^*, \theta^*))$, then by Slutsky's theorem we have the desired result.

Using the Skorokhod construction define a joint probability space for λ_n such that $\lambda_n \overset{\text{a.s.}}{\to} \theta^*$. We will condition on the sequence λ_n and use a triangular array central limit theorem for the random variables

$$g_{ni} = \frac{z_{ni}}{\bar{a}(\lambda_n)}\left(y_i - \frac{e^{(\bar{\theta}(\lambda_n)-\lambda_n)'x_i}}{1+e^{(\bar{\theta}(\lambda_n)-\lambda_n)'x_i}}\right)x_i \tag{94}$$

$$= \frac{z_{ni}}{\bar{a}(\lambda_n)}\nabla_\theta \ell(\theta-\lambda_n; x_i, y_i)\Big|_{\theta=\bar{\theta}(\lambda_n)}. \tag{95}$$

Because λ_n is independent of the data, $\mathbb{E}(f(g_{ni})|\lambda_n, z_{ni}=1) = \mathbb{E}_{\lambda_n}(f(g_{ni}))$ for any f. The triangular array CLT applies since

$$\mathbb{E}(g_{ni}|\lambda_n) = 0, \tag{96}$$

$$\operatorname{Var}(g_{ni}|\lambda_n) = \mathbb{E}\big[\operatorname{Var}(g_{ni}|\lambda_n, z_{ni})|\lambda_n\big] \tag{97}$$

$$= \mathbb{P}(z_{ni}=1|\lambda_n)\bar{a}(\lambda_n)^{-2}\operatorname{Var}_{\lambda_n}\big(\nabla_\theta \ell\big(\bar{\theta}(\lambda_n)-\lambda_n; x_{ni}, y_{ni}\big)\big) \tag{98}$$

$$= \bar{a}(\lambda_n)^{-1} J\big(\bar{\theta}(\lambda_n), \lambda_n\big) \tag{99}$$

$$\overset{\text{a.s.}}{\to} \bar{a}\big(\theta^*\big)^{-1} J\big(\theta^*, \theta^*\big). \tag{100}$$

Therefore, defining $S_n = n^{-1/2}\sum_{i=1}^n g_{ni}$ and $Z = N(0, a(\theta^*)^{-1}J(\theta^*,\theta^*))$, the CLT tells us $\mathbb{P}(S_n \in A|\lambda_n) \to \mathbb{P}(Z \in A)$ whenever $\lambda_n \to \theta^*$, which we assumed occurs with probability 1. By dominated convergence, we also have $\mathbb{P}(S_n \in A) \to \mathbb{P}(Z \in A)$.

Next we turn to the Hessian. We have $\hat{\theta}_n \overset{p}{\to} \theta^*$ by Theorem 5, so $\phi_n \overset{p}{\to} \theta^*$ as well. Writing

$$h_i^{\theta,\lambda} = \frac{e^{(\theta-\lambda)'x_i}}{(1+e^{(\theta-\lambda)'x_i})^2}x_i x_i' z_i^\lambda \tag{101}$$

we need to show that

$$\bar{a}(\lambda_n)^{-1}\big(\mathbb{E}_n h^{\phi_n,\lambda_n}\big)^{-1} \overset{p}{\to} \bar{a}\big(\theta^*\big)^{-1}\big(\mathbb{E}h^{\theta^*,\theta^*}\big)^{-1}. \tag{102}$$

Note that $\|h_i^{\theta,\lambda}\|_F \le \|x_i\|^2$, which is integrable; hence $\mathbb{E}_n h^{\theta^*,\theta^*} \overset{p}{\to} \mathbb{E}h^{\theta^*,\theta^*} = H(\theta^*,\theta^*) \succ 0$. Since $\bar{a}$ is continuous and strictly positive, and $\lambda_n \overset{p}{\to} \theta^*$, it suffices to show that

$$\big\|\mathbb{E}_n h^{\phi_n,\lambda_n} - \mathbb{E}_n h^{\theta^*,\theta^*}\big\|_F \overset{p}{\to} 0. \tag{103}$$

Note that $h_i^{\theta^*,\theta^*} = \frac{1}{4}x_i x_i'$, and define $w_{ni} = \frac{e^{(\phi_n-\lambda_n)'x_i}}{(1+e^{(\phi_n-\lambda_n)'x_i})^2}$.

Following the structure of the proof of Proposition 3, take K large enough that $\mathbb{E}\|x\|^2\mathbf{1}_{\|x\|>K} < \varepsilon$ and truncate the h_i:

$$\begin{aligned}
&\|\mathbb{E}_n h^{\phi_n,\lambda_n} - \mathbb{E}_n h^{\theta^*,\theta^*}\|_F \\
&\quad\le \|\mathbb{E}_n(h^{\phi_n,\lambda_n} - h^{\theta^*,\theta^*})\mathbf{1}_{\|x\|\le K}\|_F \qquad (104)\\
&\qquad + \|\mathbb{E}_n(h^{\phi_n,\lambda_n} - h^{\theta^*,\theta^*})\mathbf{1}_{\|x\|>K}\|_F \\
&\quad\le K^2\mathbb{E}_n|w_n z_n^{\lambda_n} - \tfrac{1}{4}z_n^{\theta^*}|\mathbf{1}_{\|x\|\le K} + 2\mathbb{E}_n\|x\|^2\mathbf{1}_{\|x\|>K}. \qquad (105)
\end{aligned}$$

The second term is eventually less than 2ε. Now, $w_{ni} - \frac{1}{4}$ is small, because

$$\left|\frac{d}{d\eta}\left(\frac{e^\eta}{(1+e^\eta)^2}\right)\right| = \left|\frac{e^\eta(e^\eta-1)}{(1+e^\eta)^3}\right| \le \frac{e^\eta}{(1+e^\eta)^2} \le \frac{1}{4}. \qquad (106)$$

Hence, by Cauchy–Schwarz

$$|w_{ni} - \tfrac{1}{4}| \le \tfrac{1}{4}\|\phi_n - \lambda_n\|\|x_i\|. \qquad (107)$$

So on the event $\{\max\|\lambda_n - \theta^*\|, \|\phi_n - \theta^*\| < \delta\}$, we have

$$\begin{aligned}
&\mathbb{E}_n|w_n z_n^{\lambda_n} - \tfrac{1}{4}z_n^{\theta^*}|\mathbf{1}_{\|x\|\le K} \qquad (108)\\
&\quad= \tfrac{1}{2}\mathbb{E}_n|(z^{\lambda_n} - z^{\theta^*})(w_n + \tfrac{1}{4}) + (z^{\lambda_n} + z^{\theta^*})(w_n - \tfrac{1}{4})|\mathbf{1}_{\|x\|\le K} \qquad (109)\\
&\quad\le \mathbb{E}_n|z^{\lambda_n} - z^{\theta^*}|\mathbf{1}_{\|x\|\le K} + \delta K. \qquad (110)
\end{aligned}$$

Finally, we can bound the first term exactly as we did in the proof of Proposition 3, defining $d_i = |z_i^{\lambda_n} - z_i^{\theta^*}|K^2\mathbf{1}_{\|x_i\|\le K}$ and $\mu_i = \mathbb{E}(d_i|x_i, y_i, \lambda_n) \le \delta K^3$. The same argument implies $\mathbb{P}(\mathbb{E}_n(d-\mu) \ge \varepsilon) \le 2\exp[-n\varepsilon^2/(2K^4)]$, so as $n\to\infty$ we have with probability approaching 1,

$$\begin{aligned}
&\|\mathbb{E}_n(h^{\phi_n,\lambda_n} - h^{\theta^*,\theta^*})\|_F \le \mathbb{E}_n(d-\mu) + \mathbb{E}_n\mu + \delta K^3 + 2\mathbb{E}_n\|x\|^2\mathbf{1}_{\|x\|>K} \qquad (111)\\
&\qquad\le 3\varepsilon + 2\delta K^3 \qquad (112)
\end{aligned}$$

so taking $\delta < \varepsilon/K^3$, the right-hand side is less than 5ε.

Acknowledgements. The authors are grateful to Jerome Friedman for suggesting that we investigate the bias of case-control sampling, and to Nike Sun for helpful comments and suggestions.

REFERENCES

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59** 19–35. MR0345332

BOTTOU, L. and BOUSQUET, O. (2008). The tradeoffs of large scale learning. *Adv. Neural Inf. Process. Syst.* **20** 161–168.

BRESLOW, N. E. and CAIN, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75** 11–20. MR0932812

BRESLOW, N. E., DAY, N. E. et al. (1980). *Statistical Methods in Cancer Research. The Analysis of Case-Control Studies* **1**. Distributed for IARC by WHO, Geneva, Switzerland.

CHAWLA, N. V., JAPKOWICZ, N. and KOTCZ, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* **6** 1–6.

FEARS, T. R. and BROWN, C. C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* **6** 955–960.

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139. MR1473055

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting. *Ann. Statist.* **28** 337–407. MR1790002

HE, H. and GARCIA, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21** 1263–1284.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

HUBER, P. J. (2011). *Robust Statistics*. Springer, Berlin.

LUMLEY, T., SHAW, P. A. and DAI, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *Int. Stat. Rev.* **79** 200–220.

MANI, I. and ZHANG, I. (2003). kNN approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*. ICML, Washington, DC.

MANSKI, C. F. and THOMPSON, T. S. (1989). Estimation of best predictors of binary response. *J. Econometrics* **40** 97–123. MR0975759

MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22** 719–748.

OWEN, A. B. (2007). Infinitely imbalanced logistic regression. *J. Mach. Learn. Res.* **8** 761–773. MR2320678

PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411. MR0556730

SCOTT, A. J. and WILD, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *J. Roy. Statist. Soc. Ser. B* **48** 170–182. MR0867995

SCOTT, A. J. and WILD, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics* **47** 497–510. MR1132540

SCOTT, A. and WILD, C. (2002). On the robustness of weighted methods for fitting models to case-control data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 207–219. MR1904701

WEBB, S., CAVERLEE, J. and PU, C. (2006). Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *Proceedings of the Third Conference on Email and Anti-Spam* (*CEAS*). CEAS, Mountain View, CA.

WEINBERG, C. R. and WACHOLDER, S. (1990). The design and analysis of case-control studies with biased sampling. *Biometrics* 963–975.

XIE, Y. and MANSKI, C. F. (1989). The logit model and response-based samples. *Sociol. Methods Res.* **17** 283–302.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
390 SERRA MALL
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: wfithian@stanford.edu
hastie@stanford.edu