A. Missing Proofs

Before we proceed to the proofs, we recall the definitions of S, E, $S_{-1,1}$, $S_{1,-1}$, $E_{-1,1}$, and $E_{1,-1}$ below:

$$S = \{x \mid f(x) \neq \widehat{f}(x)\}, \;\; E = \{x \mid f(\Delta_f(x)) \neq f(\Delta_{\widehat{f}}(x))\},$$

$$S_{-1,1} = \{x \mid f(x) = -1,\ \widehat{f}(x) = 1\}, \;\; S_{1,-1} = \{x \mid f(x) = 1, /\widehat{f}(x) = -1\},$$

$$E_{-1,1} = \{x \mid \Delta_{\widehat{f}}(x) \in S_{-1,1},\ f(\Delta_f(x)) = 1\}, \;\; E_{1,-1} = \{x \mid \Delta_f(x) \in S_{1,-1} \setminus \{x\},\ \Delta_{\widehat{f}}(x) = x\}\ .$$

A.1. Proofs of Theorems in Section 3.2

Theorem 2. *The price of opacity is strictly positive, i.e.,* $\mathrm{POP} > 0$*, if the following condition holds:*

$$\mathbb{P}_{x\sim D}\{x \in E\} > 2\mathtt{err}(f^\star, f^\star) + 2\epsilon_1. \tag{6}$$

Proof of Theorem 2. From Eq. (5) and Lemma 1, we have $\mathbb{P}_{x\sim D}\{x \in E\} = \mathrm{POP} + 2\mathrm{POP}^-$. Notice that points in E^- contribute to the error of f, implying $\mathrm{POP}^- \leq \mathtt{err}(f, f)$. Recall that $\mathtt{err}(f, f) = \mathrm{err}(f^\star, f^\star) + \epsilon_1$. Since $\mathrm{POP} + 2\mathrm{POP}^- > 2(\mathtt{err}(f^\star, f^\star) + \epsilon_1)$, we have $\mathrm{POP} > 0$, completing the proof. □

Theorem 3. *The enlargement set can be partitioned as* $E = E_{-1,1} \uplus E_{1,-1}$.

Proof. First, recall $S_{-1,1} = \{x \mid f(x) = -1 \wedge \widehat{f}(x) = 1\}$, and $S_{1,-1} = \{x \mid f(x) = 1 \wedge \widehat{f}(x) = -1\}$. Further,

$$E_{-1,1} = \{x \mid \Delta_{\widehat{f}}(x) \in S_{-1,1},\ f(\Delta_f(x)) = 1\};$$
$$E_{1,-1} = \{x \mid \Delta_f(x) \in S_{1,-1} \setminus \{x\},\ \Delta_{\widehat{f}}(x) = x\}.$$

Suppose $y = \Delta_f(x)$, and $z = \Delta_{\widehat{f}}(x)$. First assuming $f(y) \neq f(z)$, we show that $x \in E$.

Case a ($y = x$ and $z \neq x$): Since $y = x$, either $f(x) = 1$ or $\{u \mid c(x,u) < 2 \text{ and } f(u) = 1\} = \emptyset$. Moreover, as $z \neq x$, $\widehat{f}(x) = -1$, and $z = \mathrm{argmin}_u\{c(x,u) \mid c(x,u) < 2 \text{ and } \widehat{f}(u) = 1\}$. We first argue in this case that $f(x) = 1$. Suppose $f(x) = -1$. Then $\{u \mid c(x,u) < 2 \text{ and } f(u) = 1\} = \emptyset$ implying $f(z) = -1$. Since $f(x) \neq f(z)$, this gives a contradiction. If $f(x) = 1$ and $f(z) = -1$ then as $\widehat{f}(z) = 1$, $z \in S_{-1,1}$ and hence $x \in E_{-1,1}$.

Case b ($y \neq x$ and $z = x$): Again as $z = x$, either $\widehat{f}(x) = 1$ or $\{u \mid c(x,u) < 2 \text{ and } \widehat{f}(u) = 1\} = \emptyset$. Also as $y \neq x$, $f(x) = -1$ and $y = \mathrm{argmin}_u\{c(x,u) \mid c(x,u) < 2 \text{ and } f(u) = 1\}$. If $\widehat{f}(x) = 1$ then as $f(x) = -1$, $x \in S_{-1,1}$. Since $f(y) = 1$ we have $x \in E_{-1,1}$. If $\widehat{f}(x) = -1$ and $\{u \mid c(x,u) < 2 \text{ and } \widehat{f}(u) = 1\} = \emptyset$ then $\widehat{f}(y) = -1$ and $f(y) = 1$ implying $x \in E_{1,-1}$.

Case c ($y \neq x$ and $z \neq x$): Hence, $y = \{u \mid c(x,u) < 2 \text{ and } f(u) = 1\}$ and $z = \mathrm{argmin}_u\{c(x,u) \mid c(x,u) < 2 \text{ and } \widehat{f}(u) = 1\}$. Since $f(y) = 1$, by assumption it follows that $f(z) = -1$. This implies $z \in S_{-1,1}$ and $x \in E_{-1,1}$.

Now we show that if $x \in E$ then $f(\Delta_f(x)) \neq f(\Delta_{\widehat{f}}(x))$. Suppose $x \in E_{-1,1}$. Then $f(\Delta_{\widehat{f}}(x)) = -1$ and $f(\Delta_f(x)) = 1$. Similarly if $x \in E_{1,-1}$ then $f(\Delta_f(x)) = 1$ and $f(\Delta_{\widehat{f}}(x)) = -1$. □

A.2. Proofs of Propositions and Corollaries in Section 3.3

Proposition 5. *The following are sufficient for* $\mathrm{POP} > 0$*:*
(a) $\Phi_\sigma(t_f) - \Phi_\sigma(t_f - t) > \ell$ *when* $t_f > t_{\widehat{f}}$
(b) $\Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(t_f - t) > \ell$ *when* $t_f < t_{\widehat{f}}$

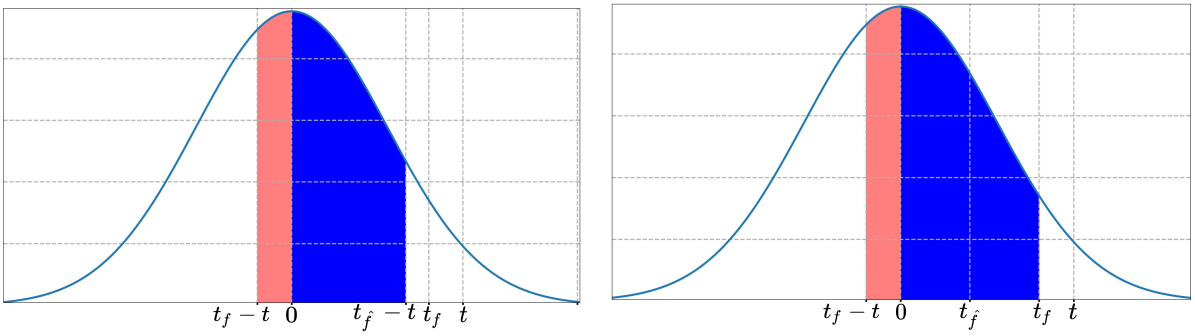

Figure 5. (**Left**) Set E when, $t_f < t_{\hat{f}}$. Blue region represents POP^+ and red region represents POP^-. (**Right**) Set E when $t_f > t_{\hat{f}}$. Blue color represents POP^+ and red color represents POP^-.

Proof. Recall $\ell = 2 \cdot \texttt{err}(f^\star, f^\star) + 2\epsilon_1$. We determine the enlargement set E in both the cases below and then in each case use Thm. 2 to conclude the proof.

Case a ($t_f > t_{\hat{f}}$): Using Theorem 3, E can be determined easily. We note E in this case in the following observation (also see Fig. 5).

Observation 10. *If $t_f > t_{\hat{f}}$ then $E = [t_f - t, t_f)$.*

From the property of normal distribution we have $\mathbb{P}_{x\sim D}\{x \in E\} = \Phi_\sigma(t_f) - \Phi_\sigma(t_f - t)$, and the result follows from Thm. 2.

Case b ($t_f < t_{\hat{f}}$): The set E in this case is as in Obs. 10 (also see Fig. 5).

Observation 11. *If $t_f < t_{\hat{f}}$ then $E = [t_f - t, t_{\hat{f}} - t)$.*

Similar to the previous case, we have $\mathbb{P}_{x\sim D}\{x \in E\} = \Phi_\sigma(t_{\hat{f}} - t) - \Phi_\sigma(t_f - t)$, and again the proof follows from Thm. 2. □

Proposition 7. *Let h be realizable and $\epsilon_2 > 2\epsilon_1$. Then, there exists $\sigma_0 > 0$ such that for all $\sigma < \sigma_0$ it holds that:*

$$\text{POP} = \begin{cases} \Phi_\sigma(t_{\hat{f}} - t) - \Phi_\sigma(|t - t_f|) & \text{if } t_f < t_{\hat{f}} \\ \Phi_\sigma(t_f) - \Phi_\sigma(|t - t_f|) & \text{if } t_f > t_{\hat{f}} \end{cases} \tag{8}$$

Proof. Since h is realizable by a threshold classifier there is an $\alpha \in \mathbb{R}$ such that $h(x) = 1$ if and only if $x \geq \alpha$. Without loss of generality, let $\alpha = 0$ (otherwise we consider distribution $\mathcal{N}(\alpha, \sigma)$). Observe that, in this case the optimal classifier $f^\star$ is defined as $f^\star(x) = 1$ for $x \geq t$ and -1 otherwise and satisfies $\texttt{err}(f^\star, f^\star) = 0$. Further, let $\epsilon_1 = \texttt{err}(f, f)$, and $\epsilon_2 = \mathbb{P}_{x\sim D}\{f(x) \neq \widehat{f}(x)\}$. We define $\sigma_0 > 0$ such that $\Phi_{\sigma_0}(\frac{t}{2}) = \frac{1}{2} + \epsilon_1$.

We show in the lemma below that $t_f > \frac{t}{2}$ for the chosen values of σ.

Lemma 12. *For all $0 < \sigma < \sigma_0$, $t_f > \frac{t}{2} > 0$.*

Proof. Suppose for contradiction $t_f \leq \frac{t}{2}$. Then $t - t_f \geq \frac{t}{2}$. This implies

$$\begin{aligned} \epsilon_1 := \texttt{err}(f, f) &= \frac{1}{2} - \Phi_\sigma(t_f - t) \\ &\underset{(i)}{=} \Phi_\sigma(t - t_f) - \frac{1}{2} \underset{(ii)}{\geq} \Phi_\sigma(\frac{t}{2}) - \frac{1}{2} \underset{(iii)}{>} \Phi_{\sigma_0}(\frac{t}{2}) - \frac{1}{2} = \epsilon_1 \end{aligned}$$

In the above, (i) follows from the symmetry property of Normal distribution, (ii) is an immediate consequence of monotonicity property of CDF function $\Phi(.)$ and finally (iii) follows from the choice of σ. This completes the proof. □

Hence, we assume throughout the proof $t_f > \frac{t}{2} > 0$. Now we determine POP and show that it greater than 0 for the two cases.

Case 1 ($t_f < t_{\widehat{f}}$): The enlargement set $E = [t_f - t, t_{\widehat{f}} - t]$ (determined in the proof of Prop. 5, see Obs. 11). Now in Obs. 13, we show that $t < t_{\widehat{f}}$ in this case.

Observation 13. *If $t_f < t_{\widehat{f}}$ then $t < t_{\widehat{f}}$.*

Proof. Note that $\epsilon_1 = \frac{1}{2} - \Phi_\sigma(t_f - t)$. Further, from definition $\epsilon_2 = \Phi_\sigma(t_{\widehat{f}}) - \Phi_\sigma(t_f)$. Suppose $t \geq t_{\widehat{f}}$. Then as $t_f > 0$

$$\Phi_\sigma(t_{\widehat{f}}) - \Phi_\sigma(t_f) \;<\; \Phi_\sigma(t - t_f) - \Phi_\sigma(0) \;=\; \epsilon_1$$

If $\epsilon_2 < \epsilon_1 < 2\epsilon_1$ then this contradicts the assumption that $\epsilon_2 \geq 2\epsilon_1$. □

Hence, either $t_f < t < t_{\widehat{f}}$ or $t < t_f < t_{\widehat{f}}$. First, we analyse the case when $t_f < t < t_{\widehat{f}}$. Recall that the price of opacity is defined as POP $= \texttt{err}(f, \widehat{f}) - \texttt{err}(f, f)$. Observe that for all $x \in [0, t_{\widehat{f}} - t)$, $h(x) = f(\Delta_f(x))$ and for all $x \in [t_f - t, 0)$, $h(x) \neq f(\Delta_f(x))$. In particular, the points in $[0, t_{\widehat{f}} - t)$ contribute positively to the price of opacity, and the points in $[t_f - t, 0)$ contribute negatively to the price of opacity. Hence,

$$\begin{aligned}
\text{POP} &= (\Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(0)) - (\Phi_\sigma(0) - \Phi_\sigma(t_f - t)) \\
&= (\Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(0)) - (\Phi_\sigma(t - t_f) - \Phi_\sigma(0)) \\
&= \Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(t - t_f)
\end{aligned}$$

Finally, in this case POP is at least 0 follows from the following observation.

Observation 14. $t_{\widehat{f}} - t > t - t_f$

Proof. By assumption we have $\epsilon_2 \geq 2\epsilon_1$. This implies that

$$\begin{aligned}
\Phi_\sigma(t_{\widehat{f}}) &\geq \Phi_\sigma(t_f) + 2(\Phi_\sigma(t - t_f) - 1/2) \\
&> \Phi_\sigma(t_f) + \Phi_\sigma(2t - t_f) - \Phi_\sigma(t_f) \\
&= \Phi_\sigma(2t - t_f)
\end{aligned}$$

The second line in the above equation follows for any $\sigma > 0$ as $0 < t_f < t < t_{\widehat{f}}$. This completes the proof. □

Similarly, when $t < t_f < t_{\widehat{f}}$, it can be shown that $E = [t_f - t, t_{\widehat{f}} - t)$, and POP $= \Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(t_f - t) > 0$.

Case 2 ($t_f > t_{\widehat{f}}$): The set $E = [t_f - t, t_f)$ in this case (see Observation 10). If $t_f > t$ then it is easy to show that POP $= \Phi_\sigma(t_f) - \Phi_\sigma(t_f - t) > 0$. Next we analyse the price of opacity for $t_f < t$. Observe that for all $x \in [0, t_f)$ $h(x) = f(\Delta_f(x)) = 1$, and for all $x \in [t_f - t, 0)$, $h(x) = -1$ and $f(\Delta_f(x)) = 1$. In particular the points in $[0, t_f)$ contribute positively to the price of opacity, whereas the points in $[t_f - t, 0)$ contribute negatively to the price of opacity (see Fig. 5). Hence,

$$\begin{aligned}
\text{POP} &= (\Phi_\sigma(t_f) - \Phi_\sigma(0)) - (\Phi_\sigma(0) - \Phi_\sigma(t_f - t)) \\
&= (\Phi_\sigma(t_f) - \Phi_\sigma(0)) - (\Phi_\sigma(t - t_f) - \Phi_\sigma(0)) \\
&= \Phi_\sigma(t_f) - \Phi_\sigma(t - t_f)\,.
\end{aligned}$$

Since $t_f > \frac{t}{2}$ (from Lem. 12), $t_f > t - t_f$, and hence, POP > 0. □

Corollary 6. *Suppose h is realizable, $t_f < t$, and $\epsilon_2 > 2\epsilon_1$. Then there exists $\sigma_0 > 0$ such that for all $\sigma < \sigma_0$,* POP > 0 *if and only if $\mathbb{P}_{x\sim D}\{E\} > 2\epsilon_1$.*

Proof. Similar to Prop. 6, without loss of generality we have $h(x) = 1$ if and only if $x \geq \alpha$. Hence, the optimal classifier $f^\star$ is defined as $f^\star(x) = 1$ for $x \geq t$ and -1 otherwise and satisfies $\texttt{err}(f^\star, f^\star) = 0$. It follows from Thm. 2 that if $\mathbb{P}_{x\sim D}\{E\} > 2\epsilon_1$ then POP > 0. Now we prove the other direction.

In Props. 5 and 6 we showed that if $t_f > t_{\widehat{f}}$ then

$$\mathbb{P}_{x\sim D}\{x \in E\} = \Phi_\sigma(t_f) - \Phi_\sigma(t_f - t)$$

$$\text{POP} = \Phi_\sigma(t_f) - \Phi_\sigma(t - t_f)\ .$$

Similarly, we also showed that if $t_f < t_{\widehat{f}}$

$$\mathbb{P}_{x\sim D}\{x \in E\} = \Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(t_f - t)$$

$$\text{POP} = \Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(t - t_f)\ .$$

From the two equations in each case we conclude the following holds in both the cases:

$$\text{POP} = \mathbb{P}_{x\sim D}\{x \in E\} + \Phi_\sigma(t_f - t) - \Phi_\sigma(t - t_f)\ .$$

Also, note that $\epsilon_1 = 1/2 - \Phi_\sigma(t_f - t)$. If POP > 0 then

$$\begin{aligned}\mathbb{P}_{x\sim D}\{x \in E\} + \Phi_\sigma(t_f - t) - \Phi_\sigma(t - t_f) &> 0\\ \mathbb{P}_{x\sim D}\{x \in E\} &> \Phi_\sigma(t - t_f) - \Phi_\sigma(t_f - t)\\ &> 1 - 2\Phi_\sigma(t_f - t) = 2\epsilon_1\end{aligned}$$

The first inequality in the last line of the above equation follows from the property of normal distribution centered at 0. □

Corollary 8. *Suppose* $\epsilon_2 > 2\sqrt{\log\frac{4}{\delta}/n}$*, where* $\delta \in (0, 1)$*. Then with probability at least* $1 - \delta$*,* POP *is as in Eq.* (7)*.*

Proof. As the strategic VC dimension of the threshold classifier for an admissible cost function is at most 2, from the strategic-PAC generalization error bound we have with probability at most δ (over the samples in T_J), $\texttt{err}(f, f) = \epsilon_1 \leq \sqrt{\frac{\log\frac{8}{\delta}}{n}}$. Using this the corollary follows from Prop. 7. □

B. Experiments

B.1. Synthetic experiments: additional results

B.1.1. Multi-Variate Normal Distribution

Our theoretical results in Sec. 3.3 consider a 1D Normal distribution. In this section we empirically study POP under a split multi-variate Normal distribution with means $\mu =$ and diagonal covariance matrix $\Sigma = 2I$. We set the cost to be linear-separable with weight vector $\alpha = (1, 0, \ldots, 0)$. We measure pop for increasing dimensionality d. Across d we fix ϵ_d by choosing $m(d)$ for which the error is at most 0.01 w.p. at least 0.95 over sample sets (i.e., as in PAC), and to simplify, mimic a setting with large n by fixing $f = f^*$ (which thresholds along the x_1 axis at 2). Results for were averaged over 50 random draws of $n = 100$ test points moved by a learned $\widehat{f}$.

Figure 6 (left, center) shows results for increasing d. As can be seen POP remains quite large even for large d. Interestingly, POP begins lower for $d = 1$ (the canonical case), then increases abruptly at $d = 2$, only to then slowly decrease until plateauing at around $d = 150$. Interestingly, the low POP at $d = 1$ is due to high variance: roughly half of the time POP ≈ 0.35, whereas the other half POP $= 0$. This behavior virtually diminishes for $d > 1$.

B.1.2. Negative POP

As noted in Sec. 3.1, POP can also be negative. Fig. 6 (right) presents such a case, in which we used a mixture of two 1D Normals centered at 0 and 1 and with diagonal unit covariances. Here, as m increases, $\widehat{f}$ better estimates f, but POP remains significantly negative throughout.

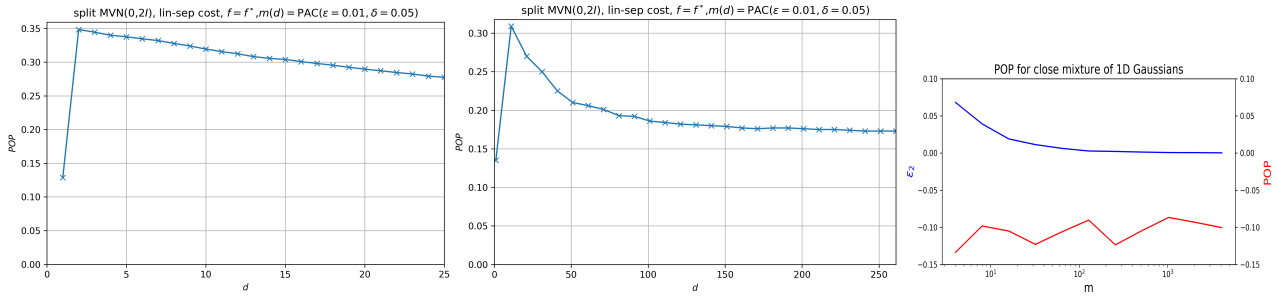

Figure 6. (**Left, center**) POP for split MVNs of increasing dimension d (center: high-resolution for small d; right: low-resolution for large d). (**Right**) Mixture of 1D Normals giving strictly negative POP.

B.2. Loans experiment: details

B.2.1. Data

Our experiments makes use of data cross-referenced from two distinct datasets related to the Prosper.com peer-to-peer loans platform. The first dataset[7] includes $n = 113,937$ examples having $d = 81$ features (not all useful for classification purposes; see below). Examples include only loan requests that have been approved, i.e., that sufficient funds were allocated by borrowers on the platform to support the request. This dataset has informative features that are useful for predictive purposes, but does not include any information on social connections. To complement this, we use a second dataset[8], which includes a full data-dump of all data available on the platform, including: loan request (both approved and disapproved), bid history, user profiles, social groups, and social connections.[9] This dataset is richer in terms of available data types, but is missing several highly informative features that appear in the first dataset. Hence, we take the intersection of both datasets (entries were matched based on user-identifying member keys). Our merged dataset includes $n = 20,222$ examples of loan requests that are described by informative features and whose corresponding users can be linked to others within the social network.

Labels. Labels in our data are determined based on a loan request's *credit grade* (`cg`), a feature indicating the risk level of the request as forecasted by the platform (for which they use proprietary tools and information that is not made public). Values of `cg` are in the range AA,A,B,C,D,E,HR, with each level corresponding to a risk value indicating the likelihood of default.[10] Because `cg` is essentially a prediction of loan default made by the platform, we use it as labels, which we binarize via $y = \mathbb{1}\{\texttt{cg} \geq B\}$ to obtain a roughly balanced dataset. In this way, learning f mimics the predictive process undertaken by the platform and for the same purposes (albeit with access to more data).

Features. Although the data includes many features, most of them are redundant for predictive purposes, others contain many missing entries, and yet others are linked directly to y in ways that make them inappropriate to use as input (e.g., borrower rate, which is fixed by the platform using a one-to-one mapping from `cg`, makes prediction trivial). Due to this, and because we also study cases in which the number of user-held examples m is small, we choose to work with a subset of six features, chosen through feature selection for informativeness and relevance. available credit, amount to loan, percent trades not delinquent, bank card utilization, total number of credit inquiries, credit history length. These remain sufficiently informative of y: as noted, a non-strategic linear baseline (on non-strategic data) achieves 84% accuracy. Fig. 2 (inlay) shows that errors in the estimation of $\widehat{f}$ (i.e., ϵ_2) are manageable even for small m.

Social network. In its earlier days, the Prosper.com platform included a social network that allowed users to form social ties beyond those derived from financial actions. In our final experiment in Sec. 4, we use the social network to determine the set of examples available for each user x for training $\widehat{f}_x$, and in particular, include in the sample set all loan requests of

[7] https://www.kaggle.com/yousuf28/prosper-loan

[8] Provided to us upon request by the authors of Kumar & Zinger (2014).

[9] The data dump is dated to 2009, shortly before Prosper.com underwent significant changes. These included changes in how loan requests are processed and attended, how credit risk is evaluated by the platform, and how users can interact. As part of these changes, Prosper.com discontinued public access to its API for obtaining data.

[10] https://www.prosper.com/invest/how-to-invest/prosper-ratings/?mod=article_inline

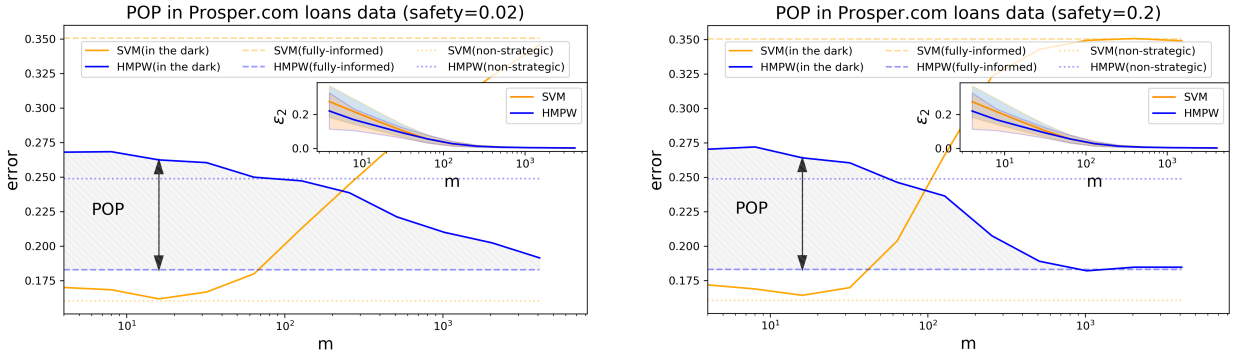

Figure 7. POP and estimation error for $\widehat{f}$ (ϵ_2) in the safe Contestant setting, for safety budgets $k = 0.02$ (1% of the total budget) and $k = 0.2$ (10% of the total budget).

any user that is at most at distance two from x (i.e., a friend or a friend-of-a-friend). The social network is in itself quite large; however, since we have informative features only for loan requests that have been approved (see above), in effect we can only make use of a smaller subset of the network which includes 994 users that have at least one social connection.

B.2.2. Experimental details

Preprocessing. All features were standardized (i.e., scaled to have zero mean and unit standard deviation).

Training. The data was split 85-15 into train and test sets, respectively. The train set was further partitioned for tuning the amount of regularization, but optimal regularization proved to be negligible, and final training was performed on the full train set. Evaluation was done on points in the held-out test set, and for settings in which $\widehat{f}$ is used, training sets for $\widehat{f}$ were sampled from the train set and labeled by f. As noted, rejection sampling was used to ensure that sample sets include at least one example from each class.

Learning algorithms. For the non-strategic baseline we used the scikit-learn implementation of SVM. For the strategic learning algorithm we used our own implementation of the algorithm of Hardt et al. (2016).

Computational infrastructure. All experiments were run on a single laptop (Intel(R) Core(TM) i7-9850H CPU 2.6GHz, 2592 Mhz, 6 core, 15.8GB RAM).

Code. Code is publicly available at `https://github.com/staretgicclfdark/strategic_rep`

B.2.3. Visualization

For visualization purposes, we embed points $x \in \mathbb{R}^6$ in $\mathbb{R}^2$. Our choice of embedding follows from the need to depict relations between the position of points in embedded space and how they are classified, i.e., their value under $f(x)$. To achieve this, we embed by partitioning features into two groups, $A_1, A_2 \subset \{1, \ldots, 6\}$, and decompose $f(x) = f_1(x) + f_2(x)$ where f_1 operates only on the subset of features in A_1 and f_2 on features in A_2. We then set the embedding ρ to be $\rho(x) = (f_1(x), f_2(x))$, i.e., values for the x-axis correspond to f_1, and values for the y-axis correspond to f_2. In this way, embedded points lie above the line $f(x) = 0$ iff they are classified as positive (note that this holds for any partitioning of features). However, such a projection does not preserve distances or directions, and so the distance of points from f and the direction and distance of point movement do not necessary faithfully represent those in the original feature space.

B.3. Loans experiment: additional results

Our main results relate to a Contestant playing $\widehat{f}$ exactly. Here we empirically study a variation on this model. If Contestant knows that $\widehat{f}$ only approximates f, she may be willing to make up for possible errors in estimation by incurring an additional cost. We refer to such a Contestant as *safe*, and model her behavior as follows. First, Contestant computes her best-response w.r.t. $\widehat{f}$, denoted x'. This provides a direction of movement $r = x' - x$. Recalling that the distance of movement was set to

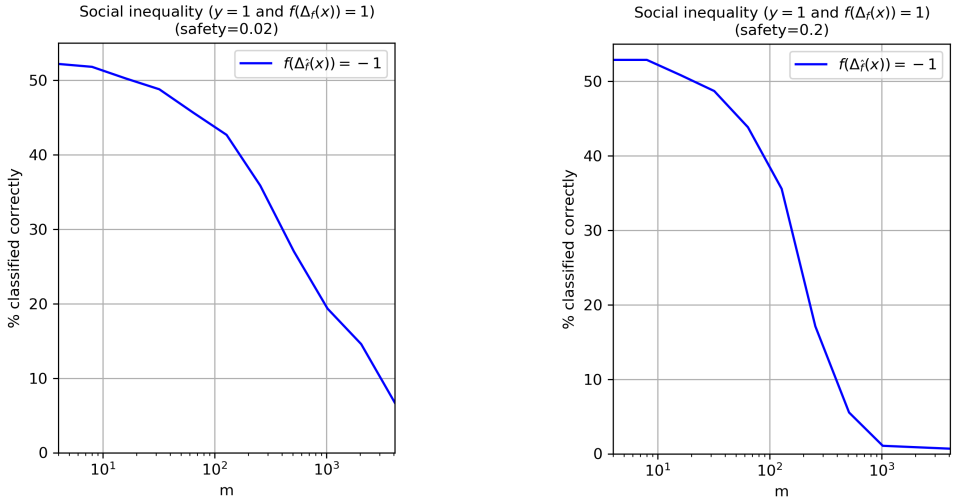

Figure 8. Social inequality for $\widehat{f}\,(\epsilon_2)$ in the safe Contestant setting, for safety budgets $k = 0.02$ (1% of the total budget) and $k = 0.2$ (10% of the total budget).

minimize the cost, the safe Contestant then invests an additional k units of cost (at most, without exceeding her overall cost budget of 2) to move further in the same direction r.

Figure 7 show POP behavior for the safe Contestant setting for 'safety' budgets $k = 0.02$ (1% of the total budget) and $k = 0.2$ (10% of the total budget). As can be seen, the main trends in the results match those in Sec. 4.2 (Fig. 2). Notice that for large m, POP is now smaller, as the increased distances can now move points beyond the region of disagreement. However, closing the POP gap requires many samples ($m \approx 1,000$) and a large safety budget (10% of total). Figure 8 shows social inequality measures for the safe Contestant setting, with matching results indicating that large m and k are necessary to maintain social equality.