A Proofs

Lemma 1. *Let $(\Omega, \mathcal{F})$ be a measurable space with a regular conditional probability property, and let $X : \Omega \to \mathbb{R}^D$, $Z : \Omega \to \mathbb{R}$ be $\mathcal{F}$-measurable random variables. Suppose P_j and P_k are σ-finite probability measures on $(\Omega, \mathcal{F})$, where P_j denotes the conditional probability measure of X given that $Z = j$, and P_k denote the same for $Z = k$, and P_j is absolutely continuous with respect to P_k. Let $f : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}$ be defined as in Section 3, and $f(x, z) \geq 0$ for all $x \in \mathbb{R}^D$, $z \in \mathbb{R}$. If the function f satisfies monotonicity in the second argument such that $f(x, j) \leq f(x, k)$ for all $x \in \mathbb{R}^D$ and for $j \leq k$, and if the Radon Nikodym derivative $\frac{dP_j}{dP_k}$ is bounded almost everywhere with respect to P_k by a finite constant $C > 0$, then*

$$E[f(X, Z)|Z = j] \leq CE[f(X, Z)|Z = k]. \quad (4)$$

Proof. Under Lemma 1's assumptions,

$$\begin{aligned} E[f(X,Z)|Z=j] &= \int_{\mathbb{R}^D} f(x,j) dP_j \\ &\leq \int_{\mathbb{R}^D} f(x,k) dP_j \\ &= \int_{\mathbb{R}^D} f(x,k) \frac{dP_j}{dP_k} dP_k \\ &\leq C \int_{\mathbb{R}^D} f(x,k) dP_k \\ &= CE[f(X,Z)|Z=k]. \end{aligned}$$

The second inequality follows from monotonicity, and the third by the Radon Nikodym theorem since $P_j << P_k$. □

Lemma 2. *Let $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^D$, $\mathcal{Z} \subseteq \mathbb{R}$. Assume that $\mathcal{X}, \mathcal{Z}$ are both finite, with $X \in \mathcal{X}$, $Z \in \mathcal{Z}$. Let $\tilde{f}$ be the projection of f onto the set of functions over $\mathcal{X} \times \mathcal{Z}$ that are monotonic with respect to Z such that for $j \leq k$, $f(x, j) \leq f(x, k)$. For $z_{(i)} \in \mathcal{Z}$, let $z_{(1)} \leq z_{(2)} \leq ... \leq z_{(|\mathcal{Z}|)}$. Define the average statistical parity violation:*

$$R_f \triangleq \sum_{i=1}^{|\mathcal{Z}|} \frac{E[f(X, Z)|Z = z_{(i)}] - E[f(X, Z)|Z = z_{(i+1)}]}{|\mathcal{Z}|}$$

Then $R_{\tilde{f}} \leq R_f$.

Proof. Let $\tilde{f} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be the projection of f onto the class of functions monotonic in the second argument, defined as follows:

$$\begin{aligned} \tilde{f} = \arg\min_{f'} & ||f - f'|| \\ \text{s.t. } & f'(x, j) \leq f'(x, k) \; \forall j, k \in \mathcal{Z}; j \leq k \end{aligned} \quad (5)$$

where

$$||f - f'||^2 = \sum_{x \in \mathcal{X}, z \in \mathcal{Z}} (f(x, z) - f'(x, z))^2.$$

The projection $\tilde{f}$ can be computed in $O(|\mathcal{X}||\mathcal{Z}|)$ time using the pool-adjacent-violators algorithm from isotonic regression (Ayer et al., 1955; JB, 1964), since a one dimensional projection can be done independently in $O(|\mathcal{Z}|)$ time for each $x \in \mathcal{X}$.

R_f is a telescoping sum:

$$R_f = \frac{E[f(X, Z)|Z = z_{(1)}] - E[f(X, Z)|Z = z_{(|\mathcal{Z}|)}]}{|\mathcal{Z}|}$$

For discrete X and Z, we have

$$E[f(X, Z)|Z = j] = \sum_{x \in \mathcal{X}} f(x, j) P(X = x|Z = j)$$

which implies

$$\begin{aligned} R_f = \frac{1}{|\mathcal{Z}|} \sum_{x \in \mathcal{X}} \Big(& f(x, z_{(1)}) P(X = x|Z = z_{(1)}) \\ & - f(x, z_{(|\mathcal{Z}|)}) P(X = x|Z = z_{(|\mathcal{Z}|)}) \Big). \end{aligned}$$

We now show that $\tilde{f}(x, z_{(1)}) \leq f(x, z_{(1)})$, and $\tilde{f}(x, z_{(|\mathcal{Z}|)}) \geq f(x, z_{(|\mathcal{Z}|)})$:
Suppose $\tilde{f}(x, z_{(1)}) > f(x, z_{(1)})$. Then we can set $\tilde{f}'(x, z_{(1)}) = f(x, z_{(1)})$ without violating the monotonicity constraints, and $||f - \tilde{f}'|| < ||f - \tilde{f}||$, which contradicts that $\tilde{f}$ solves (5). A similar argument can be made for $z_{(|\mathcal{Z}|)}$.

Since $\tilde{f}(x, z_{(1)}) \leq f(x, z_{(1)})$ and
$\tilde{f}(x, z_{(|\mathcal{Z}|)}) \geq f(x, z_{(|\mathcal{Z}|)})$, we have

$$\begin{aligned} & f(x, z_{(1)}) P(X = x|Z = z_{(1)}) \\ & \quad - f(x, z_{(|\mathcal{Z}|)}) P(X = x|Z = z_{(|\mathcal{Z}|)}) \\ \geq \; & \tilde{f}(x, z_{(1)}) P(X = x|Z = z_{(1)}) \\ & \quad - \tilde{f}(x, z_{(|\mathcal{Z}|)}) P(X = x|Z = z_{(|\mathcal{Z}|)}) \end{aligned}$$

Since the above inequality is true for all x, it holds for the sum over $x \in \mathcal{X}$, therefore $R_{\tilde{f}} \leq R_f$. □

Lemma 3. *Suppose X is a continuous (or with a straightforward extension, discrete) random variable, and let $\mathcal{S}$ be a nonempty set such that for all $x \in \mathcal{S}$, the joint probability density values $p_{X,\hat{Y}|Z=z}(x, 1) > 0$ for $z = j, k$. Suppose we have monotonicity where $f(x, j) \leq f(x, k)$ for $j \leq k$ for all $x \in \mathcal{S}$. For a binary classifier this implies $P(\hat{Y} = 1|X = x, Z = j) \leq$*

$P(\hat{Y}=1|X=x,Z=k)$. *Then we can bound one-sided statistical parity as follows:*

$$\frac{P(\hat{Y}=1|Z=j)}{P(\hat{Y}=1|Z=k)} \leq \inf_{x\in\mathcal{S}} \frac{p_{X|Z=j}(x)p_{X|\hat{Y}=1,Z=k}(x)}{p_{X|Z=k}(x)p_{X|\hat{Y}=1,Z=j}(x)}$$

Proof. Fix $x \in \mathcal{S}$. By Bayes' theorem and monotonicity,

$$\begin{aligned}
&P(\hat{Y}=1|Z=j)\\
&\quad= P(\hat{Y}=1|X=x,Z=j)\frac{p_{X|Z=j}(x)}{p_{X|\hat{Y}=1,Z=j}(x)}\\
&\quad\leq P(\hat{Y}=1|X=x,Z=k)\frac{p_{X|Z=j}(x)}{p_{X|\hat{Y}=1,Z=j}(x)}\\
&\quad= P(\hat{Y}=1|Z=k)\frac{p_{X|\hat{Y}=1,Z=k}(x)}{p_{X|Z=k}(x)}\frac{p_{X|Z=j}(x)}{p_{X|\hat{Y}=1,Z=j}(x)}
\end{aligned}$$

Since the inequality holds for all $x \in \mathcal{S}$, the tightest bound holds for the infimum. □

Lemma 4. *Let $Y \in \{0,1\}$ be a random variable representing the target. Let $\mathcal{S}$ be a nonempty set such that for all $x \in \mathcal{S}$, the following joint probability density values are non-zero for $z = j, k$: $p_{X,Y,\hat{Y}|Z=z}(x,1,1) > 0$ and $p_{X,Y|\hat{Y}=1,Z=z}(x,1) > 0$. Then,*

$$\frac{P(\hat{Y}=1|Y=1,Z=j)}{P(\hat{Y}=1|Y=1,Z=k)} \leq \inf_{x\in\mathcal{S}} \frac{c_j(x)}{c_k(x)}$$

$$\textit{where } c_z(x) = \frac{p_{X|Z=z}(x)P(Y=1|\hat{Y}=1,Z=z)}{p_{X|\hat{Y}=1,Z=z}(x)P(Y=1|Z=z)}$$

Proof. Let $\mathcal{S}$ be a nonempty set such that for all $x \in \mathcal{S}$, the following joint probability density values are non-zero for $z = j, k$:

$p_{X,Y,\hat{Y}|Z=z}(x,1,1) > 0$ and $p_{X,Y|\hat{Y}=1,Z=z}(x,1) > 0$

Fix $x \in \mathcal{S}$.

Suppose we have a monotonic binary classifier, where $P(\hat{Y}=1|X=x,Z=j) \leq P(\hat{Y}=1|X=x,Z=k)$ for $j \leq k$.

By Bayes' theorem, we have

$P(Y=1|Z=j)P(\hat{Y}=1|Y=1,Z=j)p_{X|Y=1,\hat{Y}=1,Z=j}(x)$
$= p_{X|Z=j}(x)P(\hat{Y}=1|X=x,Z=j)P(Y=1|X=x,\hat{Y}=1,Z=j)$

and $p_{X|\hat{Y}=1,Z=j}(x)P(Y=1|X=x,\hat{Y}=1,Z=j)$
$= p_{X|Y=1,\hat{Y}=1,Z=j}(x)P(Y=1|\hat{Y}=1,Z=j)$

Let $c_z(x) = \frac{p_{X|Z=z}(x)P(Y=1|\hat{Y}=1,Z=z)}{p_{X|Y=1,Z=z}(x)P(Y=1|Z=z)}$. This is well defined for $x \in \mathcal{S}$.

Combining both applications of Bayes' theorem and the monotonicity assumption:

$$\begin{aligned}
P(\hat{Y}=1 \quad &|Y=1,Z=j)\\
&= \frac{p_{X|Z=j}(x)P(Y=1|X=x,\hat{Y}=1,Z=j)}{P(Y=1|Z=j)p_{X|Y=1,\hat{Y}=1,Z=j}(x)}\\
&\quad * P(\hat{Y}=1|X=x,Z=j)\\
&= \frac{p_{X|Z=j}(x)P(Y=1|\hat{Y}=1,Z=j)}{P(Y=1|Z=j)p_{X|\hat{Y}=1,Z=j}(x)}\\
&\quad * P(\hat{Y}=1|X=x,Z=j)\\
&= c_j(x)P(\hat{Y}=1|X=x,Z=j)\\
&\leq c_j(x)P(\hat{Y}=1|X=x,Z=k)\\
&= \frac{c_j(x)}{c_k(x)}P(\hat{Y}=1|Y=1,Z=k)
\end{aligned}$$

Since this holds for all $x \in \mathcal{S}$, it holds for the infimum. □

B Counterexamples

To supplement Section 7, we give various counterexamples showing that certain relations between *statistical parity* and monotonicity do not hold.

B.1 Monotonicity does not imply statistical parity.

We show that monotonic function f may violate *one-sided statistical parity* by an example that illustrates Simpson's paradox. Suppose $X \in \{0,1\}$, where $X = 1$ means a law student passed the bar and $X = 0$ means the student did not. Let $Z \in \{0,1,2,3\}$ be the poverty level of the student, where $Z = 3$ represents the highest poverty level. Suppose $f(X,Z)$, or the admissions score, is monotonic in Z and takes the values shown in Fig. 6. Suppose that the distributions $P(X = x|Z = z)$ are given by figure 7. Then the maximum *one-sided statistical parity* violation is

$$\begin{aligned}
&E[f(X,Z)|Z=1] - E[f(X,Z)|Z=2]\\
&= f(0,1)P(X=0|Z=1) - f(0,2)P(X=0|Z=2)\\
&\quad - f(1,1)P(X=0|Z=1) + f(1,2)P(X=0|Z=2)\\
&= 1.5(0.9) - 1.5(0.1)\\
&= 1.2.
\end{aligned}$$

Thus, there is a positive one-sided satistical parity violation even though $f(X,Z)$ is monotonic in Z. This violation comes from the fact that even though $f(0,1) \leq f(0,2)$, this is outweighed by the fact that $P(X=0|Z=1) \geq P(X=1|Z=2)$. This illustrates that for a monotonic function, the *statistical parity* violation depends on the conditional probabilities $P(X = x|Z = z)$, and indeed Lemma 1 bounds

the one-sided statistcal parity violation by a ratio of conditional probabilities.

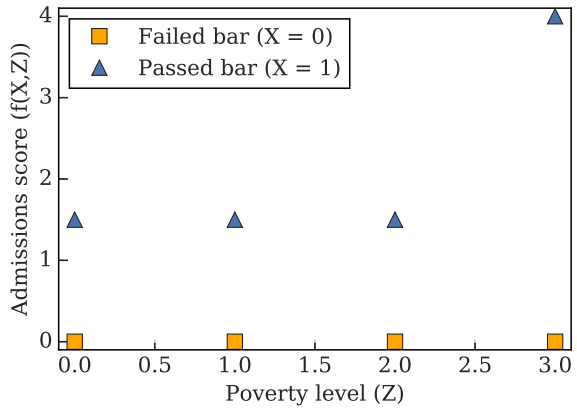

Figure 6: Monotonic admissions scores for Counterexamples B.1 and B.3.

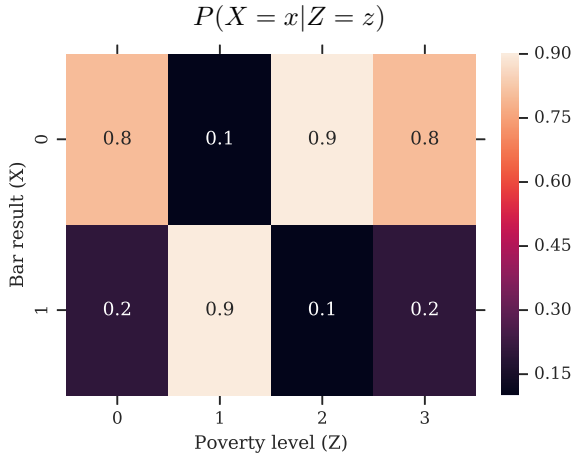

Figure 7: Distribution of X, Z for Counterexamples B.1 and B.3. The displayed values are $P(X = x|Z = z)$ for $X \in \{0, 1\}$ and $Z \in \{0, 1, 2, 3\}$.

B.2 Statistical Parity does not imply a bound on monotonicity violations.

We show that the converse of Lemma 1 does not hold: a model that satisfies *statistical parity* may have arbitrarily high monotonicity violations regardless of the likeihood ratio C. Suppose the distribution of men and women for a given height x is equal for all heights, such that $C = 1$. Suppose that *statistical parity* is satisfied such that men and women were equally likely to be selected for a sports team on average.*Statistical parity* could hold if the model accepted all men over some height h that splits the population in half (say $h = 5'8''$), and accepted all women under height h. But then for a height less than h, $P(\hat{Y} = 1|Z = \text{female}) = 0$ while $P(\hat{Y} = 1|Z = \text{male}) = 1$, and for height over h, $P(\hat{Y} = 1|Z = \text{male}) = 0$ while $P(\hat{Y} = 1|G = \text{female}) = 1$. Therefore, neither a positive nor a negative monotonicity constraint holds: there is no constant $C' > 0$ such that $P(\hat{Y} = 1|X = x, Z = \text{male}) \leq C'P(\hat{Y} = 1|X = x, Z = \text{female})$ or $P(\hat{Y} = 1|X = x, Z = \text{female}) \geq C'P(\hat{Y} = 1|X = x, Z = \text{male})$ for all x.

B.3 Monotonic projection can be more unfair in the worst case.

While Lemma 2 shows that projecting a function onto monotonicity constraints cannot increase the *average one-sided statistical parity* violation, it can increase violations *in the worst case*. Consider a continuation of the example from B.1, but this time let $f(X, Z)$ be defined by Fig. 8, and let $\tilde{f}(X, Z)$ be defined by Fig. 6. In this case, Fig. 6 is the monotonic projection of Fig. 8. Then the worst case *statistical parity* violation for the monotonic projection $\tilde{f}$ is *higher* than the worst case *statistical parity* violation for the non-monotonic f:

$$\begin{aligned} &E[\tilde{f}(X, Z)|Z = 1] - E[\tilde{f}(X, Z)|Z = 2] \\ &= 1.5(0.9) - 1.5(0.9) \\ &= 1.2 \end{aligned}$$

$$\begin{aligned} &E[f(X, Z)|Z = 1] - E[f(X, Z)|Z = 2] \\ &= 1.0(0.9) - 0.5(0.9) \\ &= 0.85 \end{aligned}$$

For a given pair j, k, as long as $\tilde{f}(x, j) \leq f(x, j)$ and $\tilde{f}(x, k) \geq f(x, k)$, then the violation

$$R_f(j, k) = E[f(X, Z)|Z = j] - E[f(X, Z)|Z = k]$$

will not be worse for the monotonic projection $\tilde{f}$: $R_{\tilde{f}}(j, k) \leq R_f(j, k)$. Lemma 2 holds because the inequalities $\tilde{f}(x, j) \leq f(x, j)$ and $\tilde{f}(x, k) \geq f(x, k)$ hold for $j = z_{(1)}$ and $k = z_{(|\mathcal{Z}|)}$, but this counterexample exists because those inequalities do not necessarily hold for any other pairs j, k in between.

C Tradeoff between likelihood ratios in Lemma 3

The bound in Lemma 3 contains two likelihood ratios: $\frac{p_{X|Z=j}(x)}{p_{X|Z=k}(x)}$ and $\frac{p_{X|\hat{Y}=1,Z=k}(x)}{p_{X|\hat{Y}=1,Z=j}(x)}$. When the first likelihood ratio is low, the second inverse likelihood ratio may be high. For example, suppose Z is an individual's poverty level (j being low poverty and k being high poverty), X is the number of extracurricular activities the individual

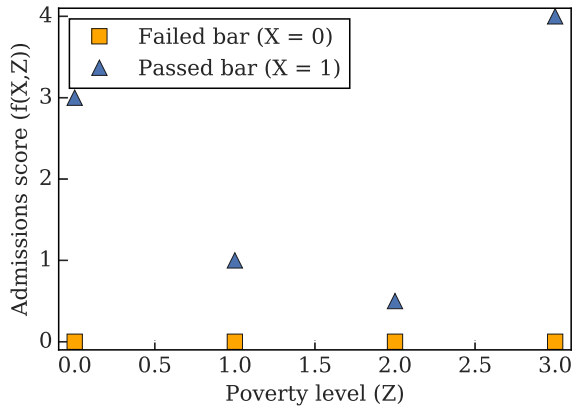

Figure 8: Nonmonotonic admissions scores for Counterexample B.3.

is involved in, and $\hat{Y} = 1$ means the individual is accepted into university. Suppose all individuals with above a certain number of extracurricular activities is accepted. Then the first likelihood ratio could be low when the number of extracurricular activities X is low. Similarly, the likelihood that a high poverty individual accepted into university has a low number of extra curricular activities is probably also higher than the likelihood that a low poverty individual accepted into university has a low number of extracurricular activities. This implies that the second inverse likelihood ratio would be high, thus trading off with the first likelihood ratio.

D Further Analysis of Law School Admissions Experiments

Figure 9 shows the distribution of the LSAT scores, undergraduate GPA, and bar exam outcomes. Examples where the bar exam outcome was missing were omitted in our experiments.

E Further Analysis of Funding Proposals Experiments

Figure 10 gives a histogram of the four different poverty levels, which are ordinal with level 3 being the most impoverished.

Figure 11 *(top)* shows the training examples' average number of exciting projects, where the error bars show the standard error of the mean. The poverty level feature ranges from 0 to 3, with 0 denoting low poverty and 3 denoting the highest poverty level. For ease of visualization, we show the quartiles of the students-reached feature.

Figure 11 *(middle)* shows the predicted probability

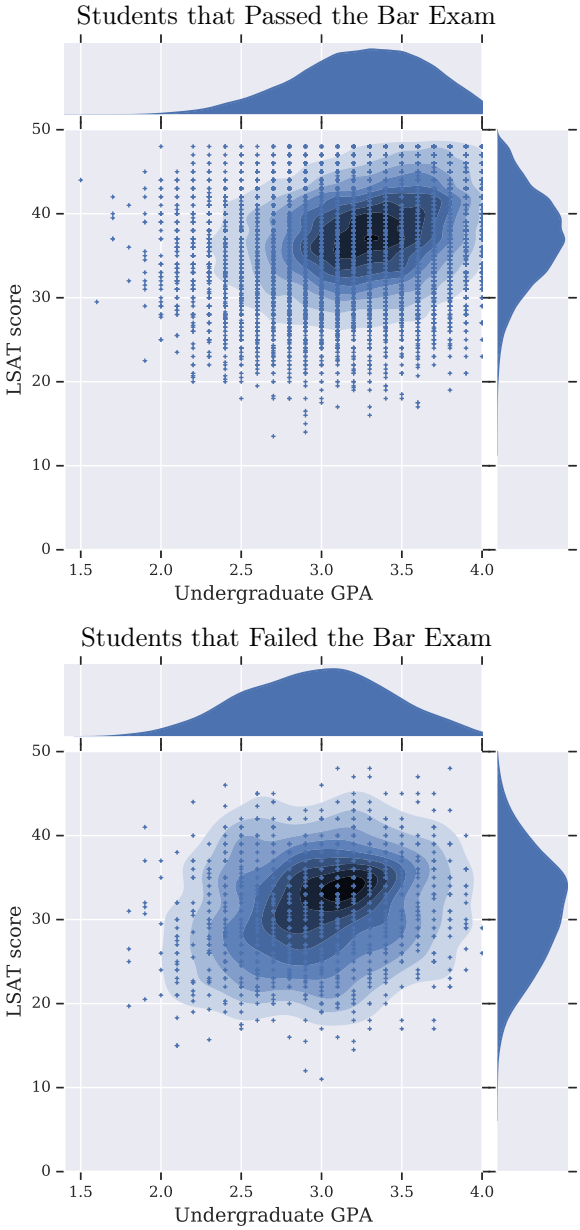

Figure 9: Distribution over the full Law School Admissions dataset of undergraduate GPA and LSAT score students for students that passed the bar exam *(top)* and students that failed the bar exam *(bottom)*. The dataset consists of 94.86% students that passed the bar exam.

that a project is exciting for a GAM model without the proposed ethical constraints. The model gives lower scores to poverty level 2 (poorer schools) than to poverty level 1 (richer schools) for every quartile of students reached. The model also gives higher scores for project that reach 30-100 students tahn to projects that reach 100+ students.

Figure 11 *(bottom)* shows that training with an ethi-

cal monotonicity shape constraint works: at the same poverty level, projects that affect more students are given a higher score. For the same quartile of students reached, the score also does not decrease for higher poverty levels.

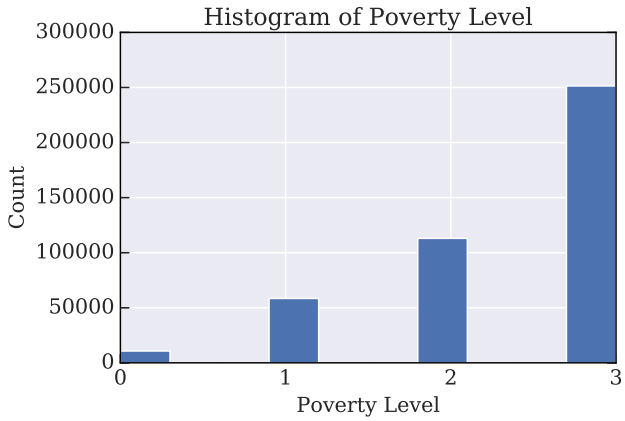

Figure 10: Histogram of the poverty level feature from the Funding Proposals dataset. 0 represents lowest poverty and 3 represents highest poverty.

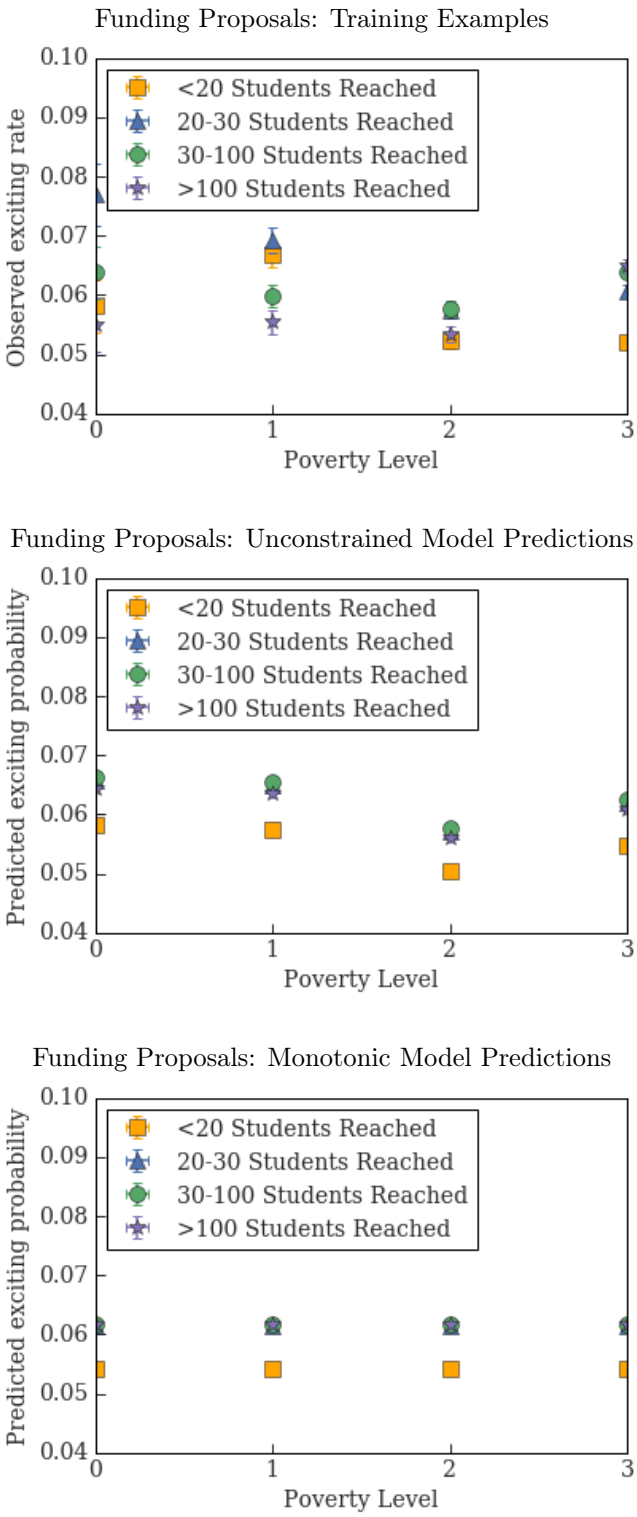

Figure 11: *(top)* Plot of the observed rate of exciting projects (mean number of exciting projects) as a function of each project's poverty level and number of students reached. Error bars show the standard deviation. *(middle)* Unconstrained model predictions. *(bottom)* Shape-constrained model predictions.