

761 A Proofs

762 A.1 Proof of Theorem 1

By the definition of $\hat{\theta}^{\pi^{(\rho)}}$, we have

$$\hat{\theta}^{\pi^{(\rho)}} = \frac{1}{n} \sum_{i=1}^n \left(f(X_i) + (Y_i - f(X_i)) \frac{\xi_i}{\pi^{(\rho)}(X_i)} \right).$$

763 From the assumption, we have $\hat{\rho} = \rho^* + o_P(1)$. By the continuity of the budget-preserving path
 764 $\pi^{(\rho)}$, it follows that $\pi^{(\hat{\rho})}(X_i) = \pi^{(\rho^*)}(X_i) + o_P(1)$ for any $i \in \{1, \dots, n\}$. This, as a result, gives
 765 $\hat{\theta}^{\pi^{(\hat{\rho})}} = \hat{\theta}^{\pi^{(\rho^*)}} + o_P(1)$ by the continuity of $\hat{\theta}^{\pi^{(\rho)}}$.

It follows from Proposition 1 in [49] that we have

$$\sqrt{n} \left(\hat{\theta}^{\pi^{(\rho^*)}} - \theta^* \right) \xrightarrow{d} \mathcal{N}(0, \sigma_{\rho^*}^2), \quad \sigma_{\rho^*}^2 = \text{Var}(\hat{\theta}^{\pi^{(\rho^*)}}).$$

Since $\hat{\theta}^{\pi^{(\hat{\rho})}} \xrightarrow{P} \hat{\theta}^{\pi^{(\rho^*)}}$,

$$\sqrt{n} \left(\hat{\theta}^{\pi^{(\hat{\rho})}} - \theta^* \right) \xrightarrow{d} \mathcal{N}(0, \sigma_{\rho^*}^2).$$

By the definition of ρ^* , $\rho^* = \arg \min_{\rho} \text{Var}(\hat{\theta}^{\pi^{(\rho)}})$, we have

$$\sigma_{\rho^*}^2 = \text{Var}(\hat{\theta}^{\pi^{(\rho^*)}}) \leq \min\{\text{Var}(\hat{\theta}^{\pi^{(0)}}), \text{Var}(\hat{\theta}^{\pi^{(1)}})\} = \min\{\sigma_0^2, \sigma_1^2\}.$$

766 This completes the proof.

767 A.2 A sufficient condition for $\hat{\rho} = \rho^* + o_P(1)$

768 **Proposition 1.** Suppose $\hat{e}^2(X) = e^2(X) + o_P(1)$, and ρ^* is unique. Suppose $\hat{e}(X)$ is uniformly
 769 upper bounded by $M > 0$. Suppose further that $\pi^{(\rho)}(X)$ is uniformly lower-bounded by $m > 0$,
 770 then we have $\hat{\rho} = \rho^* + o_P(1)$.

771 *Proof.* Denote

$$\mathcal{F} = \left\{ f_{\rho}(x) = \frac{\hat{e}^2(x)}{\pi^{(\rho)}(x)} : \rho \in [0, 1] \right\}.$$

772 We first show that \mathcal{F} is a P-Glivenko-Cantelli class.

773 Since $\pi^{(\rho)}$ is continuous, and supported on $[0, 1]$, it is uniformly continuous on $[0, 1]$. Hence for any
 774 $\delta > 0$, there exists $\eta > 0$ such that $|\pi^{(\rho_1)}(X) - \pi^{(\rho_2)}(X)| \leq \frac{m^2}{M^2} \delta$ whenever $|\rho_1 - \rho_2| \leq \eta$. Now,
 775 we cover $[0, 1]$ with a grid $0 = \rho_0 < \rho_1 < \dots < \rho_K = 1$, where $\rho_k - \rho_{k-1} = \eta$ for $k \leq K - 1$.
 776 Then, for any $\rho \in [\rho_{k-1}, \rho_k]$, we have

$$|f_{\rho}(x) - f_{\rho_{k-1}}(x)| = \left| \frac{\hat{e}^2(x)}{\pi^{(\rho)}(x)} - \frac{\hat{e}^2(x)}{\pi^{(\rho_{k-1})}(x)} \right| \leq \frac{M^2}{m^2} \left| \pi^{(\rho)}(x) - \pi^{(\rho_{k-1})}(x) \right| \leq \delta.$$

777 Hence $[f_{\rho_{k-1}} - \delta, f_{\rho_{k-1}} + \delta]$ is a 2δ -bracket in $L_1(P)$ that contains every f_{ρ} with $\rho \in [\rho_{k-1}, \rho_k]$.
 778 So the bracketing number $N_{[]}^{\cdot}(2\delta, \mathcal{F}, L_1(P)) \leq K \leq \frac{1}{\eta} + 1 < \infty$. We thus conclude
 779 from the Blum-DeHardt theorem that \mathcal{F} is a P-Glivenko-Cantelli class. Consequently, we have

$$\sup_{\rho \in [0, 1]} \left| \frac{1}{n} \sum_{i=1}^n \frac{\hat{e}^2(X_i)}{\pi^{(\rho)}(X_i)} - \mathbb{E} \frac{\hat{e}^2(X)}{\pi^{(\rho)}(X)} \right| \xrightarrow{P} 0.$$

780 This implies that

$$\left| \inf_{\rho \in [0, 1]} \frac{1}{n} \sum_{i=1}^n \frac{\hat{e}^2(X_i)}{\pi^{(\rho)}(X_i)} - \inf_{\rho \in [0, 1]} \mathbb{E} \frac{\hat{e}^2(X)}{\pi^{(\rho)}(X)} \right| \xrightarrow{P} 0.$$

781 By definition, $\hat{\rho} = \arg \min_{\rho} \frac{1}{n} \sum_{i=1}^n \frac{\hat{e}^2(X_i)}{\pi^{(\rho)}(X_i)}$. Denote $S = \arg \min_{\rho} \mathbb{E} \frac{\hat{e}^2(X)}{\pi^{(\rho)}(X)}$. Then, by continuity
 782 of $\pi^{(\rho)}(X)$, we have $d(\hat{\rho}, S) \xrightarrow{P} 0$, for $d(\hat{\rho}, S) = \inf\{|\hat{\rho} - \hat{\rho}^*| : \hat{\rho}^* \in S\}$.
 783 Now, for any $\hat{\rho}^* \in S$, we have

$$\begin{aligned} \mathbb{E} \frac{e^2(X)}{\pi^{(\hat{\rho}^*)}(X)} &\leq \mathbb{E} \frac{\hat{e}^2(X) + o_P(1)}{\pi^{(\hat{\rho}^*)}(X)} \\ &\leq \mathbb{E} \frac{\hat{e}^2(X)}{\pi^{(\rho^*)}(X)} + o_P(1) \mathbb{E} \frac{1}{\pi^{(\hat{\rho}^*)}(X)} \\ &\leq \mathbb{E} \frac{e^2(X) + o_P(1)}{\pi^{(\rho^*)}(X)} + o_P(1) \mathbb{E} \frac{1}{\pi^{(\hat{\rho}^*)}(X)} \\ &= \mathbb{E} \frac{e^2(X)}{\pi^{(\rho^*)}(X)} + o_P(1) \mathbb{E} \left[\frac{1}{\pi^{(\rho^*)}(X)} + \frac{1}{\pi^{(\hat{\rho}^*)}(X)} \right] \\ &= \mathbb{E} \frac{e^2(X)}{\pi^{(\rho^*)}(X)} + o_P(1). \end{aligned}$$

784 Since

$$\text{Var}(\hat{\theta}^{\pi^{(\rho)}}) = \mathbb{E} \left(\frac{e^2(X)}{\pi^{(\rho)}(X)} \right) + C,$$

785 where C is a constant independent of ρ , we have

$$\text{Var}(\hat{\theta}^{\pi^{(\hat{\rho}^*)}}) \leq \text{Var}(\hat{\theta}^{\pi^{(\rho^*)}}) + o_P(1).$$

786 On the other hand, by the definition of ρ^* ,

$$\text{Var}(\hat{\theta}^{\pi^{(\hat{\rho}^*)}}) \geq \text{Var}(\hat{\theta}^{\pi^{(\rho^*)}})$$

787 also holds. Whence $\text{Var}(\hat{\theta}^{\pi^{(\hat{\rho}^*)}}) \xrightarrow{P} \text{Var}(\hat{\theta}^{\pi^{(\rho^*)}})$.

788 Since ρ^* is the unique minimizer of $\text{Var}(\hat{\theta}^{\pi^{(\rho)}})$, $\hat{\rho}^* \xrightarrow{P} \rho^*$ by continuity. Since $d(\hat{\rho}, S) \xrightarrow{P} 0$ and $\hat{\rho}^*$
 789 is an arbitrary element in S , we immediately conclude that

$$\hat{\rho} \xrightarrow{P} \rho^*.$$

790

□

791 A.3 Proof of Theorem 2

By the definition of $\hat{\theta}^{\pi^{(\rho)}}$, we have

$$\hat{\theta}^{\pi^{(\rho)}} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \left(\ell_{\theta,i}^f + \left(\ell_{\theta,i} - \ell_{\theta,i}^f \right) \frac{\xi_i}{\pi^{(\rho)}(X_i)} \right).$$

792 We assume $\hat{\rho} = \rho^* + o_P(1)$. By the continuity of the budget-preserving path $\pi^{(\rho)}$, it follows that
 793 $\pi^{(\hat{\rho})}(X_i) = \pi^{(\rho^*)}(X_i) + o_P(1)$ for any $i \in \{1, \dots, n\}$. This, as a result, gives $\hat{\theta}^{\pi^{(\hat{\rho})}} = \hat{\theta}^{\pi^{(\rho^*)}} + o_P(1)$
 794 by the continuity of $\ell_{\theta,i}^f + \left(\ell_{\theta,i} - \ell_{\theta,i}^f \right) \frac{\xi_i}{\pi^{(\rho)}(X_i)}$ with respect to θ .

Given the assumption that $\hat{\theta}^{\pi^{(\rho^*)}} \xrightarrow{P} \theta^*$, from Theorem 1 in [49], we have

$$\sqrt{n} \left(\hat{\theta}^{\pi^{(\rho^*)}} - \theta^* \right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\rho^*}),$$

795 where $\Sigma_{\rho^*} = H_{\theta^*}^{-1} \text{Var} \left(\nabla \ell_{\theta^*,i}^f + \left(\nabla \ell_{\theta^*,i} - \nabla \ell_{\theta^*,i}^f \right) \frac{\xi_i}{\pi^{(\rho^*)}(X_i)} \right) H_{\theta^*}^{-1}$.

Since $\hat{\theta}^{\pi^{(\hat{\rho})}} \xrightarrow{P} \hat{\theta}^{\pi^{(\rho^*)}}$,

$$\sqrt{n} \left(\hat{\theta}^{\pi^{(\hat{\rho})}} - \theta^* \right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\rho^*}).$$

796 The definition $\rho^* = \arg \min_{\rho} \Sigma_{jj}^{\pi^{(\rho)}}$ yields

$$\Sigma_{\rho^*, jj} = \Sigma_{jj}^{\pi^{(\rho^*)}} \leq \min\{\Sigma_{jj}^{\pi^{(0)}}, \Sigma_{jj}^{\pi^{(1)}}\} = \min\{\Sigma_{0, jj}, \Sigma_{1, jj}\}.$$

797 This completes the proof.

798 B A natural family of budget-preserving paths

799 Among the diverse set of possible paths [26, 36], it is natural to consider *geodesic paths*, which are
800 a family of “shortest paths.”

801 **Definition 2** (Geodesic [7]). A curve $\gamma : I \rightarrow M$ from an interval $I \subseteq \mathbb{R}$ to a metric space M with
802 metric d is a geodesic if there is a constant $v \geq 0$ such that for any $\rho \in I$ there is a neighborhood J
803 of ρ in I such that for any $\rho_1, \rho_2 \in J$ we have

$$d(\gamma(\rho_1), \gamma(\rho_2)) = v |\rho_1 - \rho_2|.$$

804 We revisit the examples from Section 2 and provide more geodesic paths.

805 In all the following examples, we assume P and Q have the same support.

806 **Example 3** (Linear path). The linear path, $\pi^{(\rho)} \propto (1 - \rho)\pi + \rho\pi^{\text{unif}}$, is the geodesic path with
807 respect to $d(P, Q) = \|P - Q\|$ with $v = \|\pi - \pi^{\text{unif}}\|$. Here, $\|\cdot\|$ is any norm.

808 **Example 4** (Geometric path). The geometric path, $\pi^{(\rho)} \propto \pi^{1-\rho}(\pi^{\text{unif}})^{\rho}$, is the geodesic path with
809 respect to $d(P, Q) = \|\log P - \log Q\|$ with $v = \|\log \pi - \log \pi^{\text{unif}}\|$. Here, \log is taken element-
810 wise.

811 **Example 5** (Hellinger path). The Hellinger path, $\pi^{(\rho)} \propto \left((1 - \rho)\sqrt{\pi} + \rho\sqrt{\pi^{\text{unif}}}\right)^2$, is the geodesic
812 path with respect to $d(P, Q) = \|\sqrt{P} - \sqrt{Q}\|$ with $v = \|\sqrt{\pi} - \sqrt{\pi^{\text{unif}}}\|$. Here, the square root is
813 taken element-wise.

814 **Note (more examples).** Some distance metrics may not have an analytical characterization for their
815 corresponding geodesic path, such as the Wasserstein and Jensen-Shannon distances. However, it is
816 computationally tractable to solve for a geodesic path numerically up to a tolerance margin for many
817 well-defined distance metrics. For example, when computing the geodesic for the Jensen-Shannon
818 distance, we can discretize the interval $[0, 1]$ into N segments so that $P_0 = P$ and $P_N = Q$,
819 and we define a series of intermediate distributions P_1, P_2, \dots, P_{N-1} . The task is then cast as an
820 optimization problem: we minimize the total path length computed as the sum of the square roots
821 of the Jensen-Shannon divergences between successive distributions, i.e., $\sum_{i=0}^{N-1} \sqrt{\text{JS}(P_i, P_{i+1})}$.
822 Here, $\text{JS}(P\|Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M)$, where $M = \frac{1}{2}(P + Q)$. This is a constrained
823 optimization problem and can be solved by standard gradient-based methods.

824 B.1 Uniqueness of ρ^*

825 In Section 2, we saw that the uniqueness of the optimal ρ^* and the consistency of \hat{e} are sufficient
826 conditions for the consistency of $\hat{\rho}$. In the case of all three budget-preserving paths from the previous
827 section, it can be easily verified by computing the second derivative of $\text{Var}(\hat{\theta}^{\pi^{(\rho)}})$ that this variance
828 is strictly convex and thus ρ^* is unique. We include the corresponding proofs for completeness.

829 **Linear path.** We have $\pi^{(\rho)}(X) = (1 - \rho)\pi(X) + \rho\frac{n_b}{n}$. The problem of minimizing $\text{Var}(\hat{\theta}^{\pi^{(\rho)}})$ is
830 equivalent to

$$\arg \min_{\rho} \mathbb{E} \left[\frac{(Y - f(X))^2}{(1 - \rho)\pi(X) + \rho\frac{n_b}{n}} \right].$$

831 Denoting $g(\rho) = \mathbb{E} \left[\frac{(Y - f(X))^2}{(1 - \rho)\pi(X) + \rho\frac{n_b}{n}} \right]$, we have

$$g'(\rho) = \mathbb{E} \left[\frac{-(Y - f(X))^2 \left(\frac{n_b}{n} - \pi(X)\right)}{\left((1 - \rho)\pi(X) + \rho\frac{n_b}{n}\right)^2} \right],$$

832 and

$$g''(\rho) = \mathbb{E} \left[\frac{2(Y - f(X))^2 \left(\frac{n_b}{n} - \pi(X)\right)^2}{((1 - \rho)\pi(X) + \rho \frac{n_b}{n})^3} \right].$$

833 Clearly, $g''(\rho) > 0$, which means that $g(\rho)$ is convex. Hence, there is a unique optimal value of ρ in
834 $[0, 1]$.

835 Notice that $g'(1) = \frac{n^2}{n_b} \mathbb{E} [(Y - f(X))^2 (\pi(X) - \frac{n_b}{n})]$. Hence, if $\mathbb{E} [(Y - f(X))^2 \pi(X)] >$
836 $\frac{n_b}{n} \mathbb{E} [(Y - f(X))^2]$, then $g'(1) > 0$, which implies that the optimal ρ lies in $[0, 1]$.

837 **Geometric path.** Consider the path $\pi^{(\rho)}(X) \propto \pi(X)^{1-\rho}(\pi^{\text{unif}})^{\rho}$; in particular, $\pi^{(\rho)}(X) =$
838 $\frac{n_b}{n} \frac{\pi(X)^{1-\rho}}{\mathbb{E}[\pi(X)^{1-\rho}]}$.

839 Similar to the last example, we denote $g(\rho) = \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi^{(\rho)}(X)} \right] = \frac{n}{n_b} \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \right] \mathbb{E} [\pi(X)^{1-\rho}]$.

840 Then, we have

$$g'(\rho) = \frac{n}{n_b} \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \log \pi(X) \right] \mathbb{E} [\pi(X)^{1-\rho}] - \frac{n}{n_b} \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \right] \mathbb{E} [\pi(X)^{1-\rho} \log \pi(X)],$$

841 and

$$\begin{aligned} g''(\rho) &= \frac{n}{n_b} \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \log^2 \pi(X) \right] \mathbb{E} [\pi(X)^{1-\rho}] + \frac{n}{n_b} \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \right] \mathbb{E} [\pi(X)^{1-\rho} \log^2 \pi(X)] \\ &\quad - 2 \frac{n}{n_b} \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \log \pi(X) \right] \mathbb{E} [\pi(X)^{1-\rho} \log \pi(X)]. \end{aligned}$$

842 Since $(Y - f(X))^2 \geq 0$, $\pi(X) > 0$, and $\log^2 \pi(X) \geq 0$, we have that

$$\begin{aligned} &\mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \log^2 \pi(X) \right] \mathbb{E} [\pi(X)^{1-\rho}] + \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \right] \mathbb{E} [\pi(X)^{1-\rho} \log^2 \pi(X)] \\ &\geq 2 \sqrt{\mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \log^2 \pi(X) \right] \mathbb{E} [\pi(X)^{1-\rho}] \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \right] \mathbb{E} [\pi(X)^{1-\rho} \log^2 \pi(X)]} \\ &= 2 \sqrt{\mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \log^2 \pi(X) \right] \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \right] \mathbb{E} [\pi(X)^{1-\rho}] \mathbb{E} [\pi(X)^{1-\rho} \log^2 \pi(X)]} \\ &\geq 2 \sqrt{\mathbb{E}^2 \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \log \pi(X) \right] \mathbb{E}^2 [\pi(X)^{1-\rho} \log \pi(X)]} \\ &= \mathbb{E} \left[\frac{(Y - f(X))^2}{\pi(X)^{1-\rho}} \log \pi(X) \right] \mathbb{E} [\pi(X)^{1-\rho} \log \pi(X)]. \end{aligned}$$

843 The last inequality follows from the Cauchy-Schwarz inequality. Therefore, we have $g''(\rho) \geq 0$.
844 Further, if $\pi(X) \neq \pi^{\text{unif}}$, the inequality is strict, which means $g(\rho)$ is convex. Thus, there is a
845 unique optimal value of ρ in $[0, 1]$.

846 **Hellinger path.** Suppose P and Q are two discrete distributions. The Hellinger distance between
847 P and Q is $H(P, Q) = \frac{1}{\sqrt{2}} \|\sqrt{P} - \sqrt{Q}\|_2$. The geodesic connecting $\pi(X)$ and $\pi^{\text{unif}} = \frac{n_b}{n}$ is:

$$\pi^{(\rho)}(X) = \left(\frac{\sin((1 - \rho)\beta)}{\sin \beta} \sqrt{\pi(X)} + \frac{\sin(\rho\beta)}{\sin \beta} \sqrt{\frac{n_b}{n}} \right)^2,$$

848 where $\beta = \arccos\left(\sum_{i=1}^n \sqrt{\frac{\pi(X_i)}{n} \cdot n_b}\right)$.

Similarly as above, minimizing the variance $\text{Var}(\hat{\theta}^{\pi^{(\rho)}})$ amounts to minimizing the function

$$g(\rho) = \mathbb{E} \left[\frac{(Y - f(X))^2}{\left(\frac{\sin((1-\rho)\beta)}{\sin \beta} \sqrt{\pi(X)} + \frac{\sin(\rho\beta)}{\sin \beta} \sqrt{\frac{n_b}{n}} \right)^2} \right]$$

849 over ρ . The derivative $g'(\rho)$ is given by

$$-2\mathbb{E} \left[(Y - f(X))^2 \left(\frac{\sin((1-\rho)\beta)}{\sin \beta} \sqrt{\pi(X)} + \frac{\sin(\rho\beta)}{\sin \beta} \sqrt{\frac{n_b}{n}} \right)^{-3} \left(-\beta \frac{\cos((1-\rho)\beta)}{\sin \beta} \sqrt{\pi(X)} + \beta \frac{\cos(\rho\beta)}{\sin \beta} \sqrt{\frac{n_b}{n}} \right) \right],$$

850 while the second $g''(\rho)$ is given by

$$\mathbb{E} \left[(Y - f(X))^2 \left(\frac{\sin((1-\rho)\beta)}{\sin \beta} \sqrt{\pi(X)} + \frac{\sin(\rho\beta)}{\sin \beta} \sqrt{\frac{n_b}{n}} \right)^{-4} \left[6 \left(-\beta \frac{\cos((1-\rho)\beta)}{\sin \beta} \sqrt{\pi(X)} + \beta \frac{\cos(\rho\beta)}{\sin \beta} \sqrt{\frac{n_b}{n}} \right)^2 + 2\beta^2 \left(\frac{\sin((1-\rho)\beta)}{\sin \beta} \sqrt{\pi(X)} + \frac{\sin(\rho\beta)}{\sin \beta} \sqrt{\frac{n_b}{n}} \right)^2 \right] \right] > 0.$$

851 Therefore, $g(\rho)$ is strictly convex, and there is a unique optimal value of ρ in $[0, 1]$.

852 C Perturbed model errors after robust optimization

853 It is natural to choose the constraint \mathcal{C} by upper bounding a norm of ϵ . Our default choice is the
 854 ℓ_2 norm, i.e. $\|\epsilon\|_2 \leq c$. The ℓ_2 norm can be roughly thought of as controlling the variance of the
 855 errors in \hat{e}^2 . In particular, imagine $\hat{e}^2(X_i)$ can be viewed as a noisy version of $e^2(X_i)$: $\hat{e}^2(X_i) =$
 856 $e^2(X_i) + \xi_i$, where the (X_i, ξ_i) pairs are i.i.d. and ξ_i have mean zero. Then, by concentration,
 857 $\|\epsilon\|_2^2 \approx \sum_i \text{Var}(\xi_i)$.

858 In Figure 8 we illustrate how robust optimization over the ℓ_2 set \mathcal{C} recovers errors $\hat{e}^2(X_i) + \epsilon_i$ that
 859 are much closer to $e^2(X_i)$ than simply using $\hat{e}^2(X_i)$.

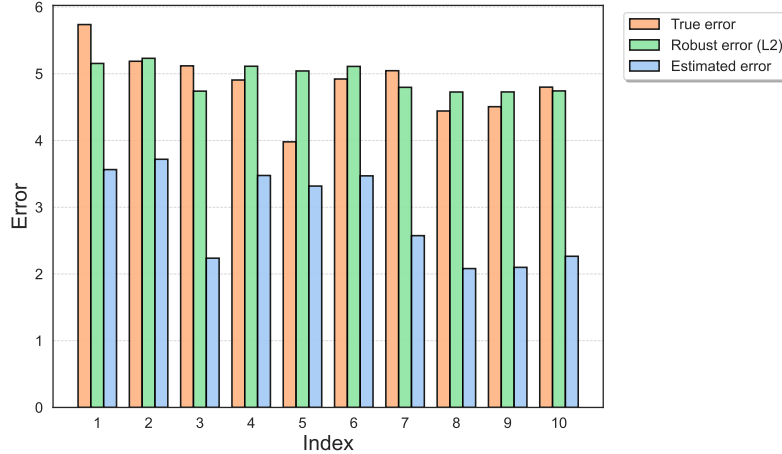


Figure 8: **Perturbed errors $\hat{e}^2(X_i) + \epsilon_i$ vs naive errors $\hat{e}^2(X_i)$ with ℓ_2 constraint \mathcal{C} .** We consider a regime where we underestimate the true error (for example, due to the model being overconfident). We let $e(X_i) \sim \mathcal{N}(5, 0.25)$ and $\hat{e}(X_i) \sim \mathcal{N}(3, 0.25)$, and $\pi^{(\rho)}$ is the linear path with $\rho = 0.5$. The robustness constraint is $\mathcal{C} = \{\epsilon : \|\epsilon\|_2 \leq 50\}$. Each index i corresponds to one sample X_i . The robust error (green bar) is the error after perturbation, $\hat{e}^2(X_i) + \epsilon_i$, and the estimated error (blue bar) is the error before perturbation, $\hat{e}^2(X_i)$. The robust errors are much closer to the estimated errors.

860 D Additional experimental results

861 In this section, we provide figures corresponding to the figures in the main text, where in addition to
 862 the effective sample size we also plot coverage.

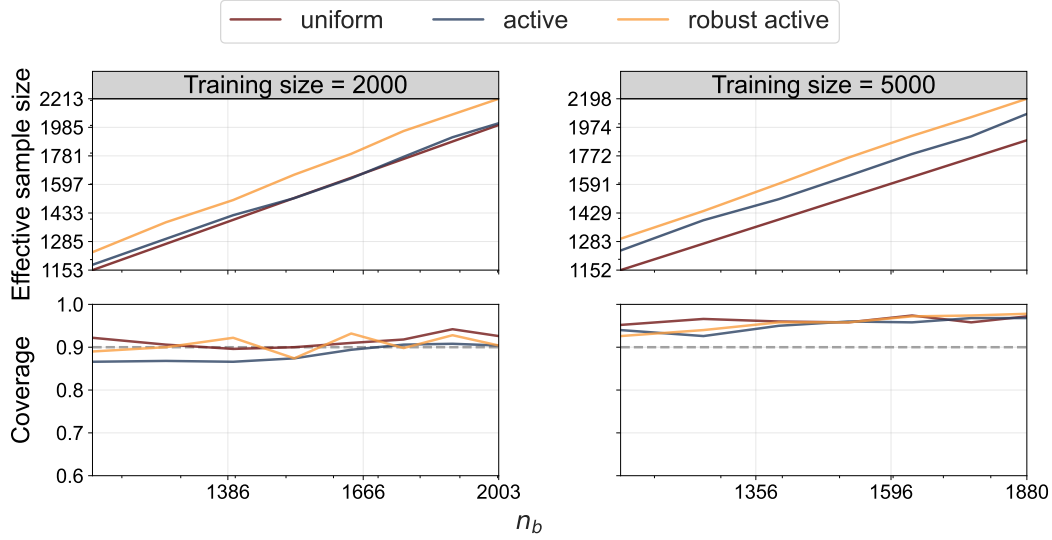


Figure 9: **Effective sample size and coverage on Pew post-election survey data**, for different dataset sizes used to train f . We compare uniform, active, and robust active sampling, for different values of the sampling budget n_b . The target of inference is the approval rate of a presidential candidate.

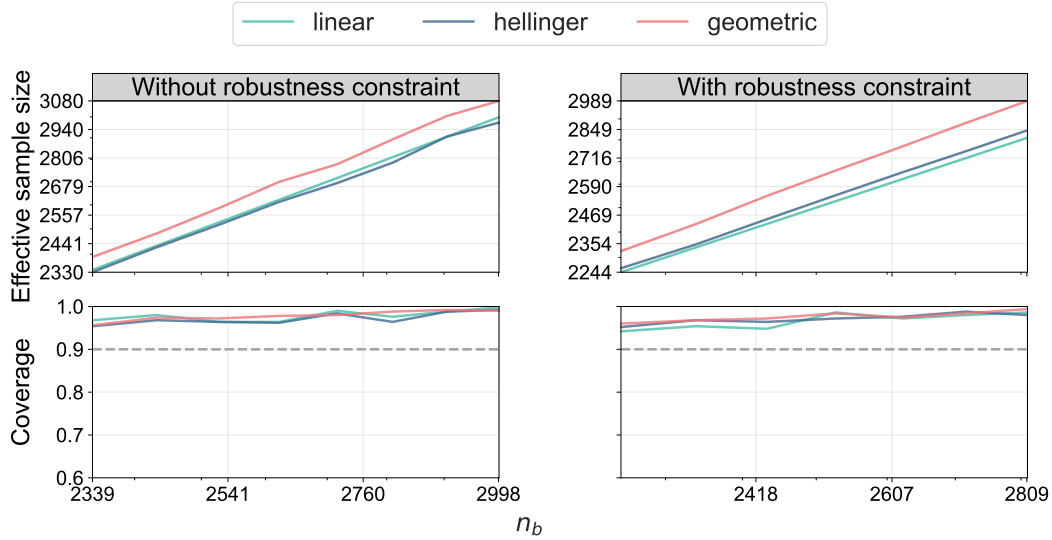


Figure 10: **Effective sample size and coverage for different budget-preserving paths on Pew post-election survey data**, without (left) and with (right) a robustness constraint \mathcal{C} . In both cases, the geometric path leads to the largest effective sample size. The target of inference is the same as in Figure 3.

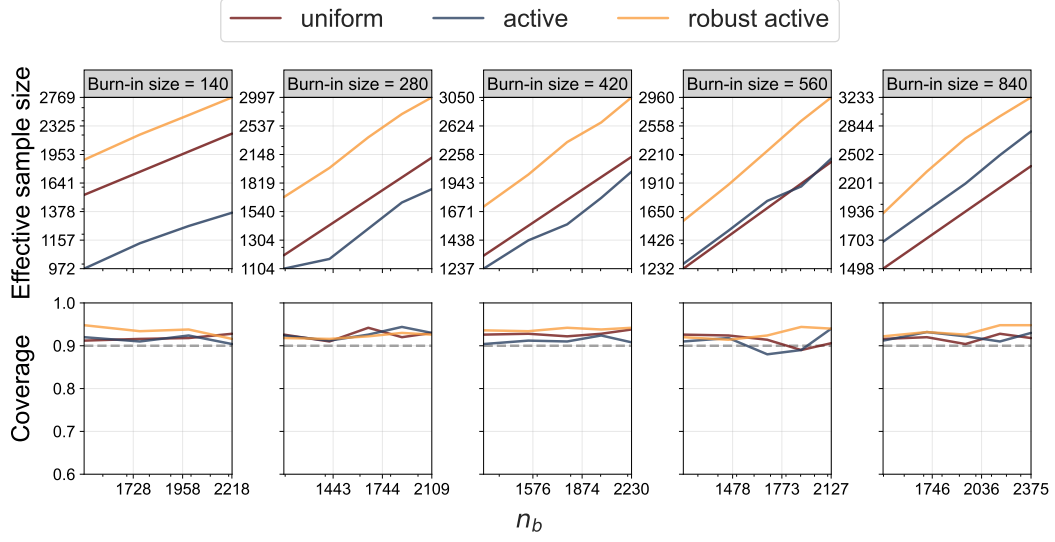


Figure 11: **Effective sample size and coverage on US Census data**, for varying burn-in dataset sizes. We compare uniform, active, and robust active sampling, for different values of the sampling budget n_b . The target of inference is the relationship between age and income, estimated via a linear regression.

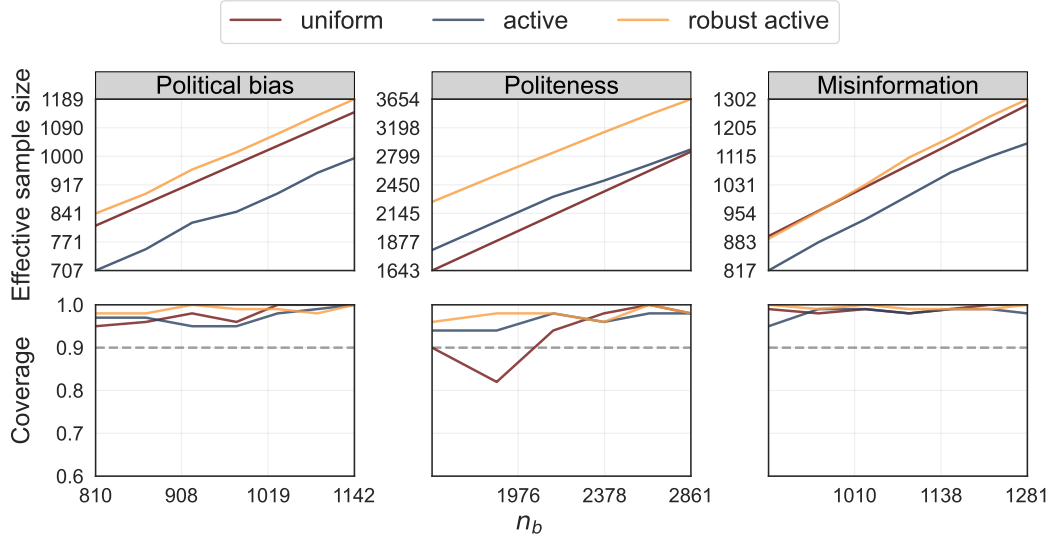


Figure 12: **Effective sample size and coverage on social science text annotation datasets**. We compare uniform, active, and robust active sampling, for different values of the sampling budget n_b . The targets of inference are (left to right) the prevalence of right-leaning political bias, the relationship between hedging and politeness, and the prevalence of misinformation.