OPTAPI: OPTIMALLY ACTIVE POWERED INFERENCE UNDER LABEL BUDGET

Anonymous authors
Paper under double-blind review

ABSTRACT

Gold-standard labeled data are essential for reliable statistical inference but are often severely limited by annotation budgets. Existing active inference methods use pre-trained predictive models to exploit unlabeled data; however, they typically focus on uniform sampling, apply no power tuning, or target only a scalar optimality criterion, leaving multivariate uncertainty unaddressed. To effectively capture joint multivariate uncertainty, we propose `OptAPI`, a two-stage optimally active powered inference framework that maximizes the utility of both labeled and unlabeled data under a label budget. The first stage deploys an active sampling rule to identify highly informative observations, while the second stage fine-tunes the prediction-powered correction. Both stages are unified under a D-optimality criterion that minimizes the determinant of the asymptotic covariance matrix, thereby directly targeting multidimensional uncertainty. Furthermore, we establish the asymptotic normality of the `OptAPI` estimator, which theoretically justifies our D-optimal design and guarantees the validity of our constructed confidence intervals. Extensive numerical evaluations on synthetic and real-world datasets demonstrate the framework's superior effectiveness and statistical efficiency.

1 INTRODUCTION

1.1 MOTIVATION AND PROBLEM FORMULATION

Gold-standard labels are essential for reliable scientific inference, but acquiring them accurately and efficiently is often expensive and time-consuming. In many modern applications, such as the expert annotation of marker genes (Baron et al., 2016; Ji et al., 2025), only a small fraction of samples can be labeled at high fidelity due to budget constraints. With the rapid development of machine learning (ML), prediction models have become more accurate and usable for subsequent analysis. More importantly, not all samples are equally informative under a fixed labeling budget. By combining ML predictions and gold-standard labels, active inference and its variants (Zrnic and Candès, 2024; Gligorić et al., 2025; Li et al., 2025; Chen et al., 2025a; Wu et al., 2026; Kluger and Bates, 2026) dynamically guide label acquisition to minimize estimation variance and conduct power tuning in the spirit of prediction-powered inference (PPI, Angelopoulos et al., 2023a;b). However, these existing active inference methods specify the sampling design without accounting for how the resulting non-uniform collection will interact with power-tuning, and the general optimally active inference framework is still largely unexplored. There arises a fundamental question:

Under a strict labeling budget, how can we best allocate a small, carefully selected set of gold-standard labels to perform valid and highly efficient inference?

To systematically answer this question, we cast the budget allocation problem through the lens of *optimal experimental design*, introducing a general framework: *Optimally Active Powered Inference under Label Budget* (`OptAPI`). Formally, let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$ be the feature and label spaces, with $(X, Y) \sim \mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$. Suppose a researcher observes a small labeled dataset $\mathcal{D}_L = \{(\widetilde{X}_k, \widetilde{Y}_k)\}_{k=1}^{n_l}$ and a large unlabeled pool $\mathcal{D} = \{X_i\}_{i=1}^{n}$ drawn i.i.d. from $\mathbb{P}_X$. Our goal is to estimate a parameter $\theta^\star \in \Theta \subset \mathbb{R}^p$ minimizing the population risk $\mathbb{E}\{\ell_\theta(X, Y)\}$ for a convex loss ℓ_θ (e.g., squared error or cross-entropy). We can only query additional labels subject to a strict budget $n_b \ll n$. Given a predictive model $f(\cdot)$, a power-tuning parameter $\lambda \geq 0$, and a sampling

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

policy $\pi(\cdot)$ satisfying the budget constraint $\mathbb{E}\{\pi(X)\} \leq n_b/n$, our active optimal estimator for the target parameter $\theta^\star$ takes the following general active inference form

$$\widehat{\theta}^{\pi}_{\lambda} = \underset{\theta\in\Theta}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \lambda \ell^{f}_{\theta,i} + \left(\ell_{\theta,i} - \lambda \ell^{f}_{\theta,i} \right) \frac{\xi_i}{\pi(X_i)} \right\} \right], \tag{1.1}$$

where $\xi_i \overset{\text{i.i.d}}{\sim} \text{Bern}(\pi(X_i))$ indicates whether the label Y_i is queried, and λ regulates the reliance on the model predictions. Our fundamental goal is to strategically allocate the labeling budget and query the possible informative labels across the unlabeled pool $\mathcal{D}$ to maximize statistical power while rigorously preserving inferential validity.

1.2 Contribution and Organization

We introduce *Optimally Active Powered Inference* (`OptAPI`), a two-stage framework that optimizes both label acquisition and prediction-powered correction under limited label budgets. By synthesizing **experimental design, active inference, and PPI**, our approach enables high-precision statistical estimation and validity through optimal power-tuning and sampling policies. See Figure 1 for an illustration of the algorithmic protocol. Our main contributions are summarized as follows.

- **General `OptAPI` framework**. We propose a D-optimal design framework that effectively leverages both labeled and unlabeled data. It flexibly accommodates tasks like mean estimation, linear regression, and general convex M-estimation via a tractable two-stage algorithm (Algorithm 1) that decouples the sampling and power-tuning phases.
- **Rigorous asymptotic guarantees**. We establish a comprehensive theoretical foundation for `OptAPI`. We characterize the D-optimal sampling policy and power-tuning parameter via the first-order conditions, compute the sampling policy with a monotone majorize-minimize algorithm, and show that the one-step τ-mixed policy never increases the asymptotic covariance determinant relative to uniform sampling. We also prove the asymptotic normality of the active estimator, which yields valid confidence intervals.
- **Extensive empirical validation**. We conduct comprehensive evaluations on both synthetic benchmarks and two real-world datasets. The results demonstrate that, by incorporating the optimal design of power-tuning and sampling policies, `OptAPI` improves estimation accuracy and effective sample size over existing state-of-the-art prediction-powered and active inference baselines while maintaining nominal coverage.

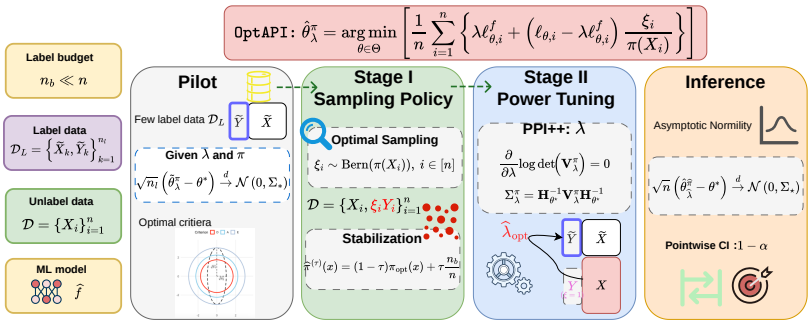


Figure 1: Protocol of optimal active inference.

Organization. Section 2 reviews related work on PPI and active inference. Section 3 details our general framework, oracle designs, practical estimation, and theoretical guarantees. Section 4 presents synthetic and real-world empirical evaluations. Section 5 concludes the paper with several future directions. The detailed proofs, additional numerical results, and extended discussions on optimality criteria, framework limitations, and computational resources are provided in Appendix.

2 EXISTING WORKS

Prediction-powered inference. Prediction-powered inference (PPI, Angelopoulos et al., 2023a) enhances the efficiency and the validity of classical inferential procedures by leveraging black-box predictions. In this framework, the researcher uses a small labeled dataset, a large unlabeled dataset, and ML predictions from a pre-trained model to construct a bias-corrected estimator for target parameters by decomposing the estimation error into two components: a model-based prediction term and a debiasing term derived from gold-standard measurements. The flourishing advances include PPI++ (Angelopoulos et al., 2023b), cross-fitting PPI (Zrnic and Candès, 2024), prediction de-correlated inference (Gan et al., 2024), stratified PPI (Fisch et al., 2024), recalibrated PPI (Ji et al., 2026), PPI on e-values (Csillag et al., 2025), MultiPPI (Cowen-Breen et al., 2026), and optimal tuning in semi-supervised learning (Shoham et al., 2025). Please refer to Ji et al. (2026); Song et al. (2026) for an overview of the PPI-related literature.

(Optimal) active statistical inference. Active statistical inference employs ML-assisted data collection to obtain trusted observations, building upon PPI and recent works on ML-based inference. Given a label budget, it maximizes estimation accuracy by proactively querying the most informative instances rather than relying on uniform sampling (Zrnic and Candès, 2024). Fundamentally, this adaptive selection process is rooted in the classical theory of optimal experimental design (Kiefer, 1959; Pukelsheim, 2006), which has been widely used in guiding optimal subsampling and active querying across massive datasets (Wang et al., 2018b;a; Yu et al., 2023; Han et al., 2025; Chen et al., 2025b). In general, ML-assisted active inference under a labeling budget can be unified by applying the augmented inverse probability weighting (AIPW) framework to our objective function. Recent literature has explored variance-minimizing active sampling rules to improve statistical efficiency (Zrnic and Candès, 2024; Ao et al., 2024; Li et al., 2025; Wu et al., 2026), with Gligorić et al. (2025) further integrating these concepts within the PPI++ framework. With different choices of power-tuning λ and sampling policy $\pi(x)$ in (1.1), we summarize the existing literature in Table 1.

Table 1: Different choices of λ and π in the general active inference framework (1.1).

Method	Power Tuning (λ)	Sampling Policy ($\pi(x)$)	Optimality
PPI (Angelopoulos et al., 2023a)	1	n_b/n (Uniform)	N/A
PPI++ (Angelopoulos et al., 2023b)	λ_{opt}	n_b/n (Uniform)	A-optimality
Active Inference (Zrnic and Candès, 2024)	1	$\propto \sqrt{\mathbb{E}\left\{\left\|\mathbf{H}^{-1}(\nabla\ell_\theta - \nabla\ell_\theta^f)\right\|_2^2 \mid X = x\right\}}$	A-optimality
Confidence-Driven Inference (Gligorić et al., 2025)	λ_{opt}	$\propto 1 - \text{Confidence}(x)$	A-optimality
Factorized Active Querying (Wu et al., 2026)	1	$\propto \sqrt{\widehat{p}(x)(1-\widehat{p}(x))}$ [(1)]	Variance-driven
Balanced Active Inference (Chen et al., 2025a)	1	$\propto \widehat{u}(x)$ via Cube Method	Covariate balance
Robust Sampling (Li et al., 2025)	1	$(1-\tau^\star)\pi_{\text{opt}}(x) + \tau^\star\pi_{\text{unif}}(x)$ [(2)]	Minimax risk
`OptAPI` (Ours)	$\boldsymbol{\lambda}_{\text{opt}}$	$\propto \sqrt{\mathbb{E}\left\{\left\|\mathbf{V}^{-1/2}(\nabla\boldsymbol{\ell}_{\boldsymbol{\theta}} - \boldsymbol{\lambda}\nabla\boldsymbol{\ell}_{\boldsymbol{\theta}}^{\boldsymbol{f}})\right\|_2^2 \mid \boldsymbol{X} = \boldsymbol{x}\right\}}$	**D-optimality**

[(1)] $\widehat{p}(x)$ is the estimated probability of a correct response for query x;
[(2)] $\tau^\star$ is optimized adversarially against mis-specified uncertainty scores.

3 METHODOLOGY

3.1 GENERAL FRAMEWORK

We consider a general convex optimization problem where the target parameter $\theta^\star \in \Theta \subset \mathbb{R}^p$ is defined as the minimizer of the population risk $\theta^\star = \arg\min_{\theta\in\Theta} \mathbb{E}\{\ell_\theta(X, Y)\}$, for some convex loss ℓ_θ. This formulation covers a broad range of statistical tasks, including mean estimation, linear regression, logistic regression, and general M-estimation. For ease of notation, we denote the instance-wise losses and their gradients as $\ell_{\theta,i} = \ell_\theta(X_i, Y_i)$, $\ell_{\theta,i}^f = \ell_\theta(X_i, f(X_i))$, $\nabla\ell_{\theta,i} = \nabla_\theta\ell_\theta(X_i, Y_i)$, and $\nabla\ell_{\theta,i}^f = \nabla_\theta\ell_\theta(X_i, f(X_i))$. The population Hessian matrix evaluated at $\theta^\star$ is denoted by $\mathbf{H}_{\theta^\star} = \mathbb{E}[\nabla^2\ell_{\theta^\star}(X, Y)]$. We design an active sampling policy $\pi\colon \mathcal{X} \to (0, 1]$ and collect label Y_i with probability $\pi(X_i)$, subject to the expected budget constraint $\mathbb{E}\{\pi(X)\} \le n_b/n$. Under the PPI++ (Angelopoulos et al., 2023b) and active inference framework with a labeling budget constraint (Zrnic and Candès, 2024), our *active optimal estimator* for $\theta^\star$ is

$$\widehat{\theta}_\lambda^\pi = \arg\min_{\theta\in\Theta} L_\lambda^\pi(\theta), \quad L_\lambda^\pi(\theta) = \frac{1}{n}\sum_{i=1}^n \left\{\lambda\ell_{\theta,i}^f + \left(\ell_{\theta,i} - \lambda\ell_{\theta,i}^f\right)\frac{\xi_i}{\pi(X_i)}\right\}, \tag{3.1}$$

where $\xi_i \overset{\text{i.i.d}}{\sim} \text{Bern}(\pi(X_i))$ is the labeling indicator for point $i \in [n]$. This construction is essentially an AIPW estimator with a particular choice of propensities $\pi(X_i)$. This formulation unifies standard PPI++ (Angelopoulos et al., 2023b) (when $\pi(x) \equiv n_b/n$) and standard active inference (Zrnic and Candès, 2024; Gligorić et al., 2025) (when $\lambda = 1$).

Lemma 1 (Unbiasedness and Asymptotic Normality)**.** *Given any fixed λ and valid sampling policy π, the active empirical risk is unbiased: $\mathbb{E}\{L^\pi_\lambda(\theta)\} = \mathbb{E}\{\ell_\theta(X,Y)\}$ for all $\theta \in \Theta$. Subsequently, under some standard regularity Assumptions 1–3, for fixed λ and π, the estimator $\widehat{\theta}^\pi_\lambda$ is consistent and asymptotically normal*

$$\sqrt{n}\left(\widehat{\theta}^\pi_\lambda - \theta^\star\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{\Sigma}^\pi_\lambda\right),$$

where $\mathbf{\Sigma}^\pi_\lambda = \mathbf{H}^{-1}_{\theta^\star}\mathbf{V}^\pi_\lambda\mathbf{H}^{-1}_{\theta^\star}$ with $\mathbf{V}^\pi_\lambda = \text{var}\left\{\lambda\nabla\ell^f_{\theta^\star,i} + \left(\nabla\ell_{\theta^\star,i} - \lambda\nabla\ell^f_{\theta^\star,i}\right)\frac{\xi_i}{\pi(X_i)}\right\}$.

Lemma 1 provides the asymptotic behavior for fixed λ and π. Sections 3.2–3.3 choose them to minimize $\mathbf{\Sigma}^\pi_\lambda$, with the design quantities for π estimated on the pilot $\mathcal{D}_L$. For many important statistical tasks, the general optimization objective in (3.1) admits an explicit closed-form solution or a straightforward estimating equation. By setting the gradient of $L^\pi_\lambda(\theta)$ to zero, the estimator $\widehat{\theta}^\pi_\lambda$ is characterized by the score equation

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\lambda\nabla\ell^f_{\theta,i} + \left(\nabla\ell_{\theta,i} - \lambda\nabla\ell^f_{\theta,i}\right)\frac{\xi_i}{\pi(X_i)}\right\} = \mathbf{0}. \tag{3.2}$$

For several foundational tasks, this admits explicit solutions or tractable Newton updates.

Mean estimation. Let $\theta^\star = \mathbb{E}(Y)$ be the population mean. Consider the squared loss $\ell_\theta(X,Y) = \frac{1}{2}||Y-\theta||_2^2$, yielding the gradient $\nabla\ell_\theta(X,Y) = \theta - Y$. Plugging $\nabla\ell_{\theta,i} = \theta - Y_i$ and $\nabla\ell^f_{\theta,i} = \theta - f(X_i)$ into the score equation (3.2) and solving yields the closed-form active mean estimator

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

Linear regression. Let $\theta^\star \in \mathbb{R}^p$ (here $p = d$) be the vector of regression coefficients minimizing the expected squared loss $\ell_\theta(X,Y) = 2^{-1}(Y - X^\top\theta)^2$, with gradient $\nabla\ell_\theta(X,Y) = -X(Y - X^\top\theta)$. Assuming the proxy model outputs continuous predictions $f(X_i) \approx Y_i$, the optimal active estimator $\widehat{\theta}^\pi_\lambda$ is the root of the linear estimating equation $\frac{1}{n}\sum_{i=1}^n X_i\left\{\lambda\big(f(X_i) - X_i^\top\theta\big) + \frac{\xi_i}{\pi(X_i)}\Big(\big(Y_i - X_i^\top\theta\big) - \lambda\big(f(X_i) - X_i^\top\theta\big)\Big)\right\} = 0$. Rearranging the terms yields a weighted ordinary least squares estimator that can be solved by

$$\widehat{\theta}^\pi_\lambda = \left\{\frac{1}{n}\sum_{i=1}^{n}\left(\lambda + (1-\lambda)\frac{\xi_i}{\pi(X_i)}\right)X_iX_i^\top\right\}^{-1}\frac{1}{n}\sum_{i=1}^{n}X_i\left\{\lambda f(X_i) + (Y_i - \lambda f(X_i))\frac{\xi_i}{\pi(X_i)}\right\}.$$

General M-estimation. For general convex M-estimation tasks (e.g., logistic regression), obtaining a closed-form solution is intractable due to the non-linearity of the gradient. Following Angelopoulos et al. (2023b) and standard semi-parametric practices, we adopt an efficient one-step quasi-Newton update to solve (3.2). Formally, given a $\sqrt{n}$-consistent initial supervised estimator $\widehat{\theta}_0$ (trained on the actively labeled data), we define the one-step active estimator

$$\widehat{\theta}^\pi_\lambda = \widehat{\theta}_0 - \widehat{\mathbf{H}}^{-1}G^\pi_\lambda(\widehat{\theta}_0), \tag{3.3}$$

where $G^\pi_\lambda(\theta) = n^{-1}\sum_{i=1}^n\{\lambda\nabla\ell^f_{\theta,i} + (\nabla\ell_{\theta,i} - \lambda\nabla\ell^f_{\theta,i})\frac{\xi_i}{\pi(X_i)}\}$ is the gradient of our active empirical loss. Here, $\widehat{\mathbf{H}}$ is a consistent estimator of the Hessian matrix, efficiently computed using only the labeled data: $\widehat{\mathbf{H}} = n^{-1}\sum_{i=1}^n\frac{\xi_i}{\pi(X_i)}\nabla^2\ell_\theta(X_i,Y_i)\big|_{\theta=\widehat{\theta}_0}$. A special case of the general M-estimation is logistic regression, where we use the cross-entropy loss $\ell_\theta(X,Y) = -Y\log\sigma(X^\top\theta) - (1-Y)\log(1-\sigma(X^\top\theta))$, the gradient terms in $G^\pi_\lambda(\widehat{\theta}_0)$ are $\nabla\ell_{\theta,i} = X_i(\sigma(X_i^\top\widehat{\theta}_0) - Y_i)$ and $\nabla\ell^f_{\theta,i} = X_i(\sigma(X_i^\top\widehat{\theta}_0) - f(X_i))$, and the Hessian is $\widehat{\mathbf{H}} = n^{-1}\sum_{i=1}^n\frac{\xi_i}{\pi(X_i)}\sigma(X_i^\top\widehat{\theta}_0)(1-\sigma(X_i^\top\widehat{\theta}_0))X_iX_i^\top$.

3.2 Oracle Optimal Design

The asymptotic normality result in Section 3.1 reduces the optimal active inference (1.1) into an optimal experimental design problem. In this Section, our objective is to find the oracle optimal sampling policy π and power-tuning parameter λ to minimize a suitable functional of the covariance matrix $\mathbf{\Sigma}_\lambda^\pi = \mathbf{H}_{\theta^\star}^{-1}\mathbf{V}_\lambda^\pi\mathbf{H}_{\theta^\star}^{-1}$ subject to the label budget constraint.

While recent active inference methods employ A-optimality (minimizing $\mathrm{tr}(\mathbf{\Sigma}_\lambda^\pi)$) (Angelopoulos et al., 2023b; Zrnic and Candès, 2024; Gligorić et al., 2025), it is scale-dependent and ignores parameter covariance, yielding suboptimal simultaneous multivariate confidence regions. To construct the tightest possible simultaneous confidence ellipsoids, we champion **D-optimality**, which minimizes the determinant $\det(\mathbf{\Sigma}_\lambda^\pi)$. Being strictly scale-invariant, it can capture the full covariance structure. Since the Hessian $\mathbf{H}_{\theta^\star}$ is independent of the design choices (π, λ), minimizing $\det(\mathbf{\Sigma}_\lambda^\pi)$ is mathematically equivalent to minimizing $\det(\mathbf{V}_\lambda^\pi)$. Unlike classical determinant-based subdata selection (e.g., IBOSS (Wang et al., 2018a)) that deterministically maximizes geometric spread in the feature space, our D-optimality operates on the asymptotic covariance of the active estimator, and the decoupled pilot design (Section 3.3) makes it computationally tractable. A rigorous geometric comparison of A- and D-optimality is provided in Appendix G and Figure A1.

Oracle active sampling (π_{opt}). Fixing λ, we optimize π. Unlike the trace, the determinant is a nonlinear functional of π, so minimizing $\det(\mathbf{V}_\lambda^\pi)$ subject to $\mathbb{E}\{\pi(X)\} \le n_b/n = \rho$ has no closed-form solution; its stationarity condition is an implicit fixed-point equation in which the optimal allocation depends on the resulting variance matrix (Theorem 1).

Theorem 1 (D-optimal sampling policy)**.** *For any fixed λ, any stationary point π_{opt} of $\det(\mathbf{V}_\lambda^\pi)$ subject to $\mathbb{E}\{\pi(X)\} \le \rho$ at which the cap $\pi \le 1$ is inactive satisfies*

$$\pi_{\mathrm{opt}}(x) \propto \sqrt{\mathbb{E}\left\{ \left\| \nabla \ell_{\theta^\star}(X,Y) - \lambda \nabla \ell^f_{\theta^\star}(X) \right\|^2_{(\mathbf{V}_\lambda^\pi)^{-1}} \,\middle|\, X = x \right\}}, \tag{3.4}$$

where the Mahalanobis norm is $\|\mathbf{z}\|^2_{\mathbf{A}} := \mathbf{z}^\top \mathbf{A}\mathbf{z}$.

Remark 1. *Theorem 1 whitens the gradient residuals by the inverse score variance $(\mathbf{V}_\lambda^\pi)^{-1}$, so the D-optimal rule favors points with both large predictive discrepancy and high statistical leverage. The A-optimal policy instead weights residuals by $\mathbf{H}_{\theta^\star}^{-2}$, $\pi^{\mathrm{A}}_{\mathrm{opt}}(x) \propto [\mathbb{E}\{\|\nabla \ell_{\theta^\star}(X,Y) - \lambda \nabla \ell^f_{\theta^\star}(X)\|^2_{\mathbf{H}_{\theta^\star}^{-2}} \mid X = x\}]^{1/2}$, ignoring the covariance structure of the prediction errors.*

Oracle power-tuning (λ_{opt}). Conversely, fixing the sampling policy π, the parameter λ dictates the degree to which the estimator trusts the machine learning predictions. Minimizing the determinant of the asymptotic score variance $\mathbf{V}_\lambda^\pi$ with respect to λ yields the following first-order characterization: $\frac{\partial}{\partial \lambda} \log\det\left(\mathbf{V}_\lambda^\pi\right) = 0$.

Theorem 2. *For any fixed π, the D-optimal power tuning λ_{opt} satisfies*

$$\lambda_{\mathrm{opt}} = \frac{\mathbb{E}\left\{ \left(\frac{1}{\pi(X)} - 1 \right) \nabla \ell^f_{\theta^\star}(X)^\top \left(\mathbf{V}^\pi_{\lambda_{\mathrm{opt}}} \right)^{-1} \nabla \ell_{\theta^\star}(X,Y) \right\}}{\mathbb{E}\left\{ \left(\frac{1}{\pi(X)} - 1 \right) \nabla \ell^f_{\theta^\star}(X)^\top \left(\mathbf{V}^\pi_{\lambda_{\mathrm{opt}}} \right)^{-1} \nabla \ell^f_{\theta^\star}(X) \right\}}.$$

Remark 2. *Theorem 2 shows that power tuning depends on the covariance geometry induced by the sampling rule. Intuitively, λ should be large when the proxy score is well aligned with the true score discrepancy and small when relying on the proxy would inflate the score variance. For completeness, the A-optimal tuning rule would instead minimize tr$(\mathbf{\Sigma}_\lambda^\pi)$, and therefore depends only on the Hessian $\mathbf{H}_{\theta^\star}$ rather than the full covariance geometry.*

3.3 Estimation

The oracle rules in Section 3.2 depend on unobserved labels and exhibit a cyclic dependency, as $\mathbf{V}_\lambda^\pi$ depends simultaneously on π and λ. To ensure empirical tractability and preserve post-sampling safety, we introduce a decoupled two-stage plugin approach. In Stage I, we fix λ to a conservative state (e.g., $\lambda = 1$) to determine and sample from the optimal policy π_{opt}. In Stage II, treating π_{opt} as

a fixed design, we optimize λ over the newly sampled data to aggressively minimize final estimation variance. We assume access to a small, pre-labeled pilot dataset $\mathcal{D}_L = \{(\widetilde{X}_k, \widetilde{Y}_k)\}_{k=1}^{n_l}$, leveraged to calibrate the design parameters before any active queries are made on the large unlabeled pool $\mathcal{D}$.

Stage I: Active sampling via pilot surrogate. Using the pilot dataset $\mathcal{D}_L$, we compute a preliminary estimate $\widetilde{\theta}_0$ and an empirical score variance matrix $\widetilde{\mathbf{V}}$, and initialize λ_0 (we take $\lambda_0 = 1$ as a conservative default). We then train an ML surrogate $u(x)$ to predict the Mahalanobis distance of gradient residuals defined by D-optimality

$$u(x) \approx \sqrt{\mathbb{E}\left\{\|\nabla \ell_{\widetilde{\theta}_0}(X,Y) - \nabla \ell^f_{\widetilde{\theta}_0}(X)\|^2_{\widetilde{\mathbf{V}}^{-1}} \mid X = x\right\}}.$$

Directly deploying this surrogate is precarious: it must be scaled to the budget n_b, and if the proxy is overconfident ($u(x) \approx 0$), inverse-probability weights explode, violating $\max_i (n\pi_i)^{-1} = O_P(1)$ and causing catastrophic variance inflation. To safeguard against these vulnerabilities, inspired by recent advances in defensive sampling and safe active learning (Zrnic and Candès, 2024; Chen et al., 2025a; Li et al., 2025), we construct a deployable policy that is simultaneously budget-constrained and minimax robust. We define the **stabilized τ-mixed sampling rule** at scale $\eta > 0$ as

$$\pi^{(\tau)}_\eta(x) = (1-\tau)\min\{\eta u(x), 1\} + \tau \frac{n_b}{n}, \tag{3.5}$$

where the scalar $\eta > 0$ serves only to match the expected number of queried labels to the label budget n_b, and $\tau \in (0,1)$ is a mixing parameter. In finite samples, we typically estimate η via

$$\widehat{\eta} = \max\left\{\eta > 0 : \sum_{i=1}^{n} \min\{\eta u(X_i), 1\} \le n_b\right\},$$

and deploy $\widehat{\pi}^{(\tau)} = \pi^{(\tau)}_{\widehat{\eta}}$. This mixture design offers two profound advantages over hard thresholding heuristics. Statistically, it guarantees a strict safety lower bound $\widehat{\pi}^{(\tau)}(X_i) \ge \tau n_b/n > 0$, which implies Assumption 1, bounding importance weights and preserving asymptotic normality even if the proxy fails. Algorithmically, convex combinations inherently preserve the exact labeling budget, avoiding unstable renormalization steps. Sensitivity analysis for τ is presented in Appendix G.5.

Remark 3 (Computing the D-optimal policy)**.** *Theorem 1 defines π_{opt} only implicitly, because $\mathbf{V}^\pi_\lambda$ depends on π itself, so we compute it via majorize-minimize (MM) algorithm. Starting from uniform sampling $\pi^{(0)} \equiv n_b/n$, each iteration $k \in [K]$ alternates two steps: (1) freeze the whitening matrix at the current variance, $\mathbf{A}_k = \mathbf{V}_\lambda^{\pi^{(k)}}$; and (2) water-fill, $\pi^{(k+1)}(x) = \min\{\eta_k u_{A_k}(x), 1\}$, where $u_A(x)^2 = \mathbb{E}\{\|\nabla \ell_{\theta^\star}(X,Y) - \lambda \nabla \ell^f_{\theta^\star}(X)\|^2_{A^{-1}} \mid X = x\}$ and η_k matches the budget. Step (2) exactly minimizes the tangent-plane upper bound of $\log\det \mathbf{V}^\pi_\lambda$ at $\mathbf{A}_k$, which is linear in $\mathbf{V}^\pi_\lambda$. Consequently, (a) each iteration never increases $\det(\mathbf{\Sigma}^\pi_\lambda)$; (b) the policy stops changing exactly at the capped policies of Theorem 1, which are first-order stationary; and (c) after one iteration, the τ-mixed policy* (3.5) *is never worse than uniform sampling for any $\tau \in [0,1)$, and further iterations cannot increase the determinant of the unmixed policy. Stage I implements this loop with the pilot plug-ins $\widetilde{\theta}_0$, $\widetilde{\mathbf{V}}$, and λ_0 ($K = 10$ iterations in our experiments); formal statements and proofs are given in Proposition C.1 of Appendix C.*

Stage II: Power-tuning on actively sampled data. After querying true labels using $\widehat{\pi}^{(\tau)}$, we must tune λ. Rather than freezing λ at a potentially unstable pilot estimate λ_0, maximizing statistical efficiency dictates dynamically optimizing λ over the newly acquired, highly informative active samples, inspired by recent semi-supervised advancements (Angelopoulos et al., 2023b; Shoham et al., 2025). Given the preliminary estimator $\widetilde{\theta}_0$, we construct the AIPW-adjusted asymptotic score variance matrix parameterized by λ

$$\widehat{\mathbf{V}}(\lambda) = \widetilde{\mathbf{V}} + \frac{1}{n}\sum_{i=1}^{n} \widehat{\Delta}_{\lambda,i}\widehat{\Delta}_{\lambda,i}^\top \frac{\xi_i}{\widehat{\pi}^{(\tau)}(X_i)}\left(\frac{1}{\widehat{\pi}^{(\tau)}(X_i)} - 1\right),$$

where $\widehat{\Delta}_{\lambda,i} = \nabla \ell_{\widetilde{\theta}_0}(X_i, Y_i) - \lambda \nabla \ell^f_{\widetilde{\theta}_0}(X_i)$, and $\widetilde{\mathbf{V}}$ is empirical variance of $\nabla \ell_{\widetilde{\theta}_0}$ evaluated over the pilot dataset $\mathcal{D}_L$ to ensure finite-sample stability. Setting the derivative of $\log\det \widehat{\mathbf{V}}(\lambda)$ to zero gives an empirical analogue of Theorem 2, but it is an implicit fixed-point equation because $\widehat{\mathbf{V}}(\cdot)$ depends on λ itself (Appendix B), and solving it iteratively is fragile in finite samples. Instead, we

exploit the fact that λ is a scalar. As established in PPI++ (Angelopoulos et al., 2023b), empirical risk minimization over a bounded parameter space enjoys uniform convergence, so reusing actively sampled data to estimate λ preserves asymptotic normality. We therefore solve the bounded problem $\widehat{\lambda}_{\text{opt}} = \arg\min_{\lambda\in[0,1]} \log\det\{\widehat{\mathbf{V}}(\lambda) + \epsilon\mathbf{I}\}$, where ϵ is a negligible penalty (e.g., 10^{-6}) ensuring strict positive definiteness, and plug $\widehat{\lambda}_{\text{opt}}$ into the final estimator $\widehat{\theta}^{\pi}_{\lambda}$.

3.4 Theoretical Guarantees

In this section, we establish the rigorous theoretical foundations of $\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}}$ for constructing asymptotic normality (Theorem 3) and confidence intervals (Theorem 4) under the general M-estimation setting. All regularity assumptions and technical proofs are deferred to Appendices A and D.

Theorem 3 (Data-dependent Asymptotic Normality)**.** *Suppose Assumptions 1–2 hold, and write π_η for the deployed policy $\pi_\eta^{(\tau)}$ of* (3.5)*. Let $(\eta^\star, \lambda^\star)$ denote fixed values such that, as in Zrnic and Candès (2024), $\mathbb{P}(\widehat{\eta} \neq \eta^\star) \to 0$, and $\widehat{\lambda} = \lambda^\star + o_p(1)$, and let $n \to \infty, n_b \to \infty$, and $n_b/n \to \rho \in (0,1)$ according to Assumption 3. Then, the data-dependent active estimator is $\sqrt{n}$-consistent and asymptotically normal*

$$\sqrt{n}(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}} - \theta^\star) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma}_\star),$$

where $\mathbf{\Sigma}_\star = \mathbf{H}_{\theta^\star}^{-1}\text{var}\left(\lambda^\star\nabla\ell^f_{\theta^\star} + (\nabla\ell_{\theta^\star} - \lambda^\star\nabla\ell^f_{\theta^\star})\frac{\xi^{\eta^\star}}{\pi_{\eta^\star}(X)}\right)\mathbf{H}_{\theta^\star}^{-1}$.

Theorem 4 (Confidence Interval)**.** *Suppose the conditions of Theorem 3 hold. Let $\widehat{\mathbf{\Sigma}}$ be a consistent plug-in estimator of the asymptotic covariance matrix such that $\widehat{\mathbf{\Sigma}} \xrightarrow{p} \mathbf{\Sigma}_\star$. For a target significance level $\alpha \in (0,1)$ and $j \in [p]$, a valid $(1-\alpha)$ confidence interval for $\theta_j^\star$ is*

$$\mathcal{C}_\alpha(\theta_j^\star) = \left[\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda},j} - z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathbf{\Sigma}}_{jj}}{n}}, \quad \widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda},j} + z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathbf{\Sigma}}_{jj}}{n}}\right], \tag{3.6}$$

where $z_{1-\alpha/2}$ is the upper $\alpha/2$-quantile of the standard normal distribution. The interval achieves nominal coverage asymptotically $\lim_{n\to\infty}\mathbb{P}\left\{\theta_j^\star \in \mathcal{C}_\alpha(\theta_j^\star)\right\} = 1-\alpha$.

Remark 4. *Theorems 3–4 give asymptotic normality and valid confidence intervals under the estimated policy and $\widehat{\lambda}$. The asymptotic variance reduces to that of active inference (Zrnic and Candès, 2024) when $\lambda^\star = 1$ and to PPI++ (Angelopoulos et al., 2023b) under uniform sampling. We summarize the two-stage procedure in Algorithm 1.*

Algorithm 1 The `OptAPI` Algorithm

Input: Labeled dataset $\mathcal{D}_L$, Unlabeled dataset $\mathcal{D}$, ML predictive model f, label budget n_b, mixture parameter $\tau \in (0,1)$, target significance level α.

1: **Pilot**: On $\mathcal{D}_L$, compute the pilot estimators $\widetilde{\theta}_0$, $\mathbf{H}_{\widetilde{\theta}_0}$, and $\widetilde{\mathbf{V}}$, and train uncertainty proxy $u(x)$.
2: **Stage I (Optimal active sampling)**: Construct the stabilized τ-mixing sampling policy $\widehat{\pi}^{(\tau)}(X_i)$ for all $X_i \in \mathcal{D}$ via (3.5) with estimated scaling parameter $\widehat{\eta}$. Then draw labeling decisions via $\xi_i \overset{\text{i.i.d}}{\sim} \text{Bern}(\widehat{\pi}^{(\tau)}(X_i))$, and further collect the label Y_i when $\xi_i = 1$ for $i \in [n]$, then construct sampled dataset $\mathcal{D} = \{X_i, \xi_i Y_i\}_{i=1}^n$.
3: **Stage II (Power-tuning)**: Estimate $\widehat{\lambda}$ by minimizing $\log\det\{\widehat{\mathbf{V}}(\lambda) + \epsilon\mathbf{I}\}$ over $\lambda \in [0,1]$.
4: **Estimation**: Compute the active empirical risk minimizer $\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}}$ via (1.1).
5: **Inference**: Construct $(1-\alpha)$- confidence intervals $\mathcal{C}_\alpha(\theta_j^\star)$ via (3.6).

Output: Return $\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}}$ and $\mathcal{C}_\alpha(\theta_j^\star)$.

4 Experiments

Baselines. We benchmark `OptAPI` against several state-of-the-art baselines: Classical Inference $(\pi_{\text{unif}}, \lambda = 0)$; PPI $(\pi_{\text{unif}}, \lambda = 1)$ (Angelopoulos et al., 2023a); PPI++ $(\pi_{\text{unif}}, \lambda^{\text{A}}_{\text{opt}})$ (Angelopoulos et al., 2023b); Active Inference $(\pi^{\text{A}}_{\text{opt}}, \lambda = 1)$ (Zrnic and Candès, 2024); and Confident Active Inference $(\pi^{\text{A}}_{\text{opt}}, \lambda^{\text{A}}_{\text{opt}})$ (Gligorić et al., 2025), serving as the strongest A-optimal competitor.

Evaluation metrics. Following Gligorić et al. (2025), we evaluate statistical efficiency using Effective Sample Size (n_{eff}), quantifying the exact number of uniformly sampled points required to achieve the active method's confidence ellipsoid volume $n_{\text{eff}} = n_b \times \{\det(\widehat{\mathbf{\Sigma}}_{\text{classical}})/\det(\widehat{\mathbf{\Sigma}}_{\text{method}})\}^{1/p}$, where $\widehat{\mathbf{\Sigma}}_{\text{classical}}$ and $\widehat{\mathbf{\Sigma}}_{\text{method}}$ denote the estimated asymptotic covariance matrices of the classical estimator and the evaluated method, respectively. For statistical validity, we also report the empirical coverage (the proportion of iterations where the true $\theta^\star$ falls within the joint 95% Wald confidence ellipsoid, i.e., $\alpha = 0.05$) and the Mean Squared Error (MSE).

General setup. For all synthetic experiments, we simulate a large unlabeled pool of $n = 10{,}000$ points. We allocate $n_l = 500$ queries to form the initial pilot dataset $\mathcal{D}_L$, which is used only to calibrate the design parameters. We sweep the active labeling budget n_b from 5% to 95% of n, averaging the results over 500 Monte Carlo replications. Unless stated otherwise, all experiments use $\tau = 0.10$, $\lambda_0 = 1$, and $K = 10$ MM iterations for the D-optimal policy; for the regression tasks, the Stage I surrogate exploits the rank-one structure of the score (Appendix G.4). All active baselines use the same pilot, surrogate class, and τ-mixing, so the methods differ only in their sampling criterion and power tuning. Due to space constraints, core results for mean estimation and linear regression are presented below; extended validations (logistic/Poisson regression), a second real-data application on census data, and ablation studies are deferred to Appendix G.

4.1 Mean Estimation under Scale Disparity

We consider estimation of the mean $\theta^\star \in \mathbb{R}^3$ under a heterogeneous Gaussian mixture designed to create a scale trap. The data consist of a majority cluster $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with probability 0.9 and a high-variance minority cluster $\mathcal{N}(\mathbf{5}, 10\mathbf{I})$ with probability 0.1, where the third coordinate is additionally inflated by a factor of 50. The proxy model $f(X)$ is systematically biased by downscaling predictions for the minority cluster by a factor of 0.5.

All methods maintain valid 95% empirical coverage across budgets (Figure 2), but with different efficiency performance. Trace-based A-optimal baselines are attracted to the inflated coordinate and can be up to 24% less efficient than the Classical baseline. By contrast, `OptAPI` uses a D-optimal criterion that balances information across directions, avoiding this scale trap. It achieves up to a 157.7% gain in n_{eff} relative to the nominal labeled budget at $n_b = 500$, and consistently outperforms uniformly sampled PPI++. These results show that active design and power tuning are both necessary for efficient multivariate inference.

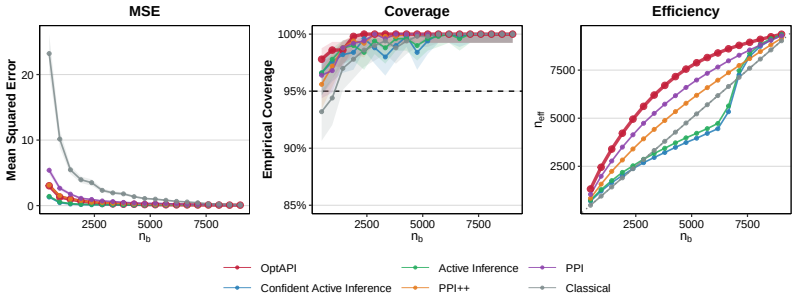


Figure 2: Performance on multivariate mean estimation.

4.2 Heteroskedastic Linear Regression

We consider a linear regression model $Y = X^\top \theta^\star + \epsilon$, where $X \in \mathbb{R}^4$ consists of an intercept and three log-normal features with scale disparities $(1, 1, 0.01)$ and extreme feature-dependent heteroskedasticity scaling as $0.5 + 2||X_{2:3}||_2^2$. We set $\theta^\star = (1.0, 2.0, -1.5, 5.0)^\top$ to pair the largest coefficient with the smallest feature scale. To simulate an imperfect proxy, we generate predictions as $f(X) \sim \mathcal{N}(X^\top \theta_{\text{biased}}, 0.5^2)$. This injects random predictive variance around a mean that is systematically misspecified by setting $\theta_{\text{biased}} = \theta^\star + c$, where $c = (0.5, -0.8, 1.2, -2.0)^\top$.

Under uniform sampling, PPI suffers from the biased proxy and performs worse than the classical baseline (Figure 3). The proposed `OptAPI` mitigates this issue by adaptively downweighting the proxy through power tuning, taking $\lambda \to 0$. The D-optimal sampling policy in `OptAPI` focuses

labeling effort on high-leverage and high-variance regions, improving efficiency under heteroskedasticity. As a result, `OptAPI` attains the largest effective sample size across label budgets.

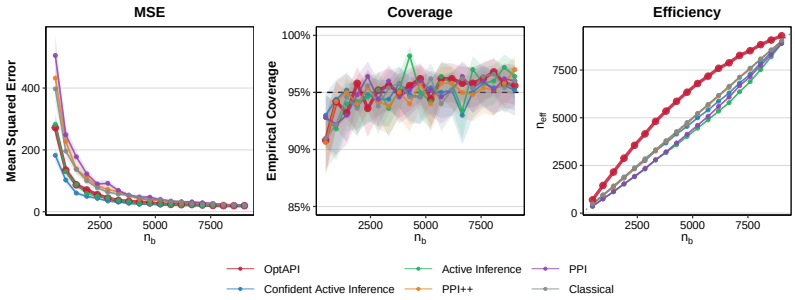


Figure 3: Performance on linear regression.

4.3 Real-World Application: California Housing Prices

Dataset. We evaluate `OptAPI` on the California Housing dataset available at `https://www.kaggle.com/datasets/camnugent/california-housing-prices/data`. We consider $n = 20{,}433$ samples and use four features (median income, housing age, total rooms, population) to estimate log-transformed median house values. We split the data into a pilot set ($n_l = 1,000$) to train a Random Forest surrogate $f(X)$, and an unlabeled pool ($n = 19,433$). Adopting a Full-Data Oracle methodology, we define the ground-truth $\theta^\star$ via Ordinary Least Squares over the full dataset, and iteratively query labels from the unlabeled pool across various budgets.

Results. The proposed `OptAPI` consistently achieves the smallest MSE relative to the Full-Data Oracle across nearly all budgets (Figure 4). Rather than being misdirected by raw marginal variances, the proposed `OptAPI` adapts to the housing feature space's structural geometry to extract maximum inferential information per label, outperforming trace-based optimal and uniform sampling baselines in overall effective sample size (n_{eff}).

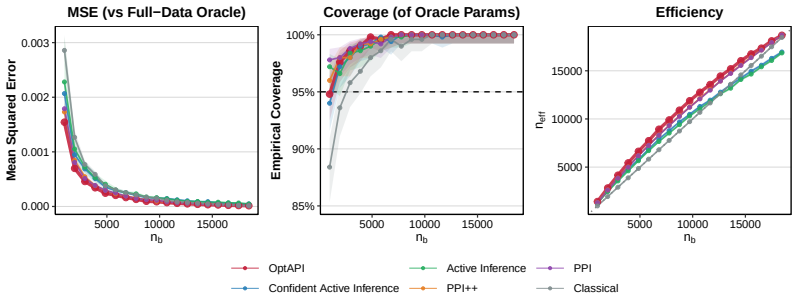


Figure 4: Real-world evaluation on the California housing prices dataset.

5 Concluding Remark

We introduced `OptAPI`, a unifying framework for budget-aware optimal active inference that integrates active inference with optimal design. By decoupling power tuning from sampling design, we construct a tractable two-stage algorithm whose sampling policy and power tuning target the D-optimality criterion, and we then establish valid inference for the resulting active estimator that attains the minimal asymptotic variance. Extensive numerical simulations and two real-data applications show that `OptAPI` improves estimation accuracy and effective sample size over existing prediction-powered and active inference baselines while maintaining nominal coverage.

Several directions merit further exploration. An essential assumption in our current setup is that the labeled and unlabeled data are drawn from the same underlying distribution. Relaxing `OptAPI` to accommodate distribution shift or covariate drift is a critical next step. Furthermore, while our current two-stage algorithm elegantly ensures optimal tractability, advancing to a jointly optimal active inference paradigm poses a profound theoretical challenge that warrants future investigation.

AI USE STATEMENT

In this work, we used generative AI tools for writing and debugging the experimental code, checking the validity of the proofs, and checking English grammar and style. All research ideas, methodological developments, theoretical results, experimental designs, and figures are our own, and every AI-drafted passage was based on content we supplied. We revised all AI-drafted text and checked it against the underlying derivations and experimental outputs. We independently verified any potential errors in the proofs flagged by AI tools and corrected them where needed. We reviewed and tested all AI-assisted code, and all reported results were implemented by running the final code. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human-subjects research, and we collected no new data. All experiments use either synthetic data or publicly available datasets: the California Housing dataset (Section 4.3) and the census healthcare benchmark of Angelopoulos et al. (2023a), derived from de-identified American Community Survey public-use records (Appendix G.4). We did not attempt to re-identify individuals or link these records to other sources. The census analysis illustrates statistical efficiency only: the estimated association between income and health insurance is not causal and should not inform decisions about individuals. More broadly, `OptAPI` is designed to reduce the risk of drawing scientific conclusions from biased machine learning predictions: its debiased estimator and power tuning keep inference valid under proxy misspecification (Lemma 1, Theorems 3–4). The τ-mixed sampling rule gives each unit a labeling probability of at least $\tau n_b/n$, so no subpopulation is excluded from the labeled sample. These guarantees rely on the pilot, pool, and target population sharing a common distribution. Under distribution shift, coverage may fail, and practitioners should check this assumption before relying on the resulting intervals (Appendix F).

REPRODUCIBILITY STATEMENT

The proposed `OptAPI` procedure is specified in Section 3.3 and summarized in Algorithm 1. All technical assumptions are stated in Appendix A. Complete proofs are in Appendix B (Theorems 1-2), Appendix C (Proposition C.1), and Appendix D (Theorems 3-4). The data generating processes, baselines, and evaluation metrics for every experiment are described in Section 4 and Appendix G. Implementation choices are detailed in Appendix G.4; these include the mixing parameter, the initial power-tuning value, the number of MM iterations, and the surrogate model. Sensitivity to these choices is examined in the ablations (Appendices G.5 -G.7). Both real datasets are publicly available, and their sources and preprocessing steps are given in Section 4.3 and Appendix G.4. The software environment and hardware are listed in Appendix G.1. The R code needed to reproduce every table and figure, with fixed random seeds, will be uploaded upon publication.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Ruicheng Ao, Hongyu Chen, and David Simchi-Levi. Prediction-guided active experiments. *arXiv preprint arXiv:2411.12036*, 2024.

Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems*, 3(4):346–360, 2016.

Boyu Chen, Zhixiang Zhou, Liuhua Peng, and Zhonglei Wang. Balanced active inference. In *Advances in Neural Information Processing Systems*, 2025a.

Haolin Chen, Holger Dette, and Jun Yu. Multi-resolution subsampling for linear classification with massive data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1260–1280, 2025b.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. *arXiv preprint arXiv:2603.27414*, 2026.

Daniel Csillag, Claudio Jose Struchiner, and Guilherme Tegoni Goedert. Prediction-powered e-values. In *Forty-second International Conference on Machine Learning*, 2025.

Adam Fisch, Joshua Maynez, R Hofer, Bhuwan Dhingra, Amir Globerson, and William W Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. *Advances in Neural Information Processing Systems*, 37:111489–111514, 2024.

Feng Gan, Wanfeng Liang, and Changliang Zou. Prediction de-correlated inference: A safe approach for post-prediction inference. *Australian & New Zealand Journal of Statistics*, 66(4): 417–440, 2024.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533, 2025.

Yixin Han, Jun Yu, Nan Zhang, Cheng Meng, Ping Ma, Wenxuan Zhong, and Changliang Zou. Leverage classifier: Another look at support vector machine. *Statistica Sinica*, 35(1):1605–1625, 2025.

Wenlong Ji, Weizhe Yuan, Emily Getzen, Kyunghyun Cho, Michael I Jordan, Song Mei, Jason E Weston, Weijie J Su, Jing Xu, and Linjun Zhang. An overview of large language models for statisticians. *arXiv preprint arXiv:2502.17814*, 2025.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of ai. *Biometrika*, page asag053, 2026.

Jack Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304, 1959.

Daniel Kluger and Stephen Bates. M-estimation under two-phase multiwave sampling with applications to prediction-powered inference. *arXiv preprint arXiv:2602.16933*, 2026.

Puheng Li, Tijana Zrnic, and Emmanuel J Candès. Robust sampling for active statistical inference. *Neural Information Processing Systems*, 2025.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

Noa Shoham, Ron Dorfman, Shalev Shaer, Kfir Y Levy, and Yaniv Romano. Prediction-powered semi-supervised learning with online power tuning. In *Advances in Neural Information Processing Systems*, 2025.

Yilin Song, Dan M Kluger, Harsh Parikh, and Tian Gu. Demystifying prediction powered inference. *arXiv preprint arXiv:2601.20819*, 2026.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405, 2018a.

HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018b.

Skyler Wu, Yash Nair, and Emmanuel J Candès. Efficient evaluation of llm performance with statistical guarantees. *arXiv preprint arXiv:2601.20251*, 2026.

Jun Yu, Jiaqi Liu, and HaiYing Wang. Information-based optimal subdata selection for non-linear models. *Statistical Papers*, 64(4):1069–1093, 2023.

Tijana Zrnic and Emmanuel J Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

Appendix for "OptAPI: Optimally Active Powered Inference under Label Budget"

Overview. Section A provides the key lemmas and some technical assumptions used in the main paper. Section B constructs the optimal designs for the power-tuning parameter and subsampling policy. Section C presents the majorize-minimize characterization of the D-optimal design and its proof. Detailed proofs of the asymptotic normality and confidence interval are presented in Section D. Section E compares the different optimality criteria. Section F provides some limitations. Section G provides the detailed experimental setup and additional empirical results.

A Useful Lemmas and Technical Assumptions

The following lemma is the asymptotic normality of the M-estimator; Please see Theorem 5.23 in Vaart (1998) for more details.

Lemma A.1. *Let Θ be an open subset of a Euclidean space, and let $x \mapsto m_\theta(x)$ be a measurable function for each $\theta \in \Theta$. Suppose that, for P-almost every x, the map $\theta \mapsto m_\theta(x)$ is differentiable at θ_0, with derivative $\dot{m}_{\theta_0}(x)$. Assume moreover that there exists a measurable envelope function m satisfying $Pm^2 < \infty$ such that, for all θ_1, θ_2 in a neighborhood of θ_0,*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le m(x)\,\|\theta_1 - \theta_2\|.$$

Assume further that the map $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion at θ_0, where θ_0 is a maximizer, with a non-singular, symmetric second-derivative matrix V_{θ_0}. If

$$P_n m_{\widehat{\theta}_n} \ge \sup_\theta P_n m_\theta - o_p(n^{-1}) \quad \textit{and} \quad \widehat{\theta}_n \xrightarrow{P} \theta_0,$$

then

$$\sqrt{n}\,(\widehat{\theta}_n - \theta_0) = -\,V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}_{\theta_0}(X_i) + o_p(1).$$

Consequently,

$$\sqrt{n}\,(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\big(0,\, V_{\theta_0}^{-1}\, P\big[\dot{m}_{\theta_0}\dot{m}_{\theta_0}^{\top}\big]\, V_{\theta_0}^{-1}\big)\,.$$

Before presenting the main theoretical results, we detail the necessary regularity conditions. Assumption 1 is a standard constraint ensuring that the sampling probabilities are well-behaved and safely bounded away from zero, preventing any single observation from dominating the inverse probability weights. Assumption 2 imposes standard smoothness and strong convexity conditions on the loss function ℓ_θ for M-estimation. These regularity conditions are present in the asymptotic proofs of Angelopoulos et al. (2023b) and Zrnic and Candès (2024) and are customary in the classical M-estimation literature (Vaart, 1998). Finally, Assumption 3 defines the asymptotic scaling of the labeling budget (n_b) relative to the total unlabeled pool (n). The limit $\rho \in (0, 1)$ corresponds to the prediction-powered and active inference regime, in which the labeling budget is a strictly positive fraction of the total data (Angelopoulos et al., 2023b; Zrnic and Candès, 2024). The vanishing-fraction regime $\rho = 0$ of optimal subsampling (Wang et al., 2018a) would require a $\sqrt{n_b}$ rather than $\sqrt{n}$ normalization and is not covered by our results.

Assumption 1 (Bounded probabilities)**.** *The sampling probabilities π_i are bounded away from zero and satisfy*

$$\max_{1 \le i \le n} (n\pi_i)^{-1} = O_p(1).$$

Assumption 2 (Smoothness)**.** *The loss ℓ is smooth if:*

- *$\ell_\theta(x, y)$ is differentiable at $\theta^\star$ for all (x, y);*
- *ℓ_θ is locally Lipschitz around $\theta^\star$: there is a neighborhood of $\theta^\star$ such that $\ell_\theta(x, y)$ is $C(x, y)$-Lipschitz and $\ell_\theta(x, f(x))$ is $C(x)$-Lipschitz in θ, where $\mathbb{E}\left\{C(X, Y)^2\right\} < \infty$, $\mathbb{E}\left\{C(X)^2\right\} < \infty$;*
- *$L(\theta) = \mathbb{E}\left\{\ell_\theta(X, Y)\right\}$ and $L^f(\theta) = \mathbb{E}\{\ell_\theta(X, f(X))\}$ have Hessian $\mathbf{H}_{\theta^\star} = \nabla^2 L(\theta^\star) \succ 0$.*

Assumption 3 (Asymptotic Labeling). *The total unlabeled pool size n and the labeling budget n_b both grow to infinity ($n \to \infty, n_b \to \infty$) such that their ratio converges to some $\rho \in (0,1)$. That is,*

$$\lim_{n, n_b \to \infty} \frac{n_b}{n} = \rho \in (0,1).$$

B Proofs for Optimal Sampling and Tuning Results

Proof of Theorem 1 Recall the asymptotic covariance is $\mathbf{\Sigma}_\star = \mathbf{H}_{\theta^\star}^{-1} \mathbf{V}_\star \mathbf{H}_{\theta^\star}^{-1}$. Using the multiplicative property of the determinant, we have

$$\det(\mathbf{\Sigma}_\star) = \det(\mathbf{H}_{\theta^\star}^{-1}) \det(\mathbf{V}_\star) \det(\mathbf{H}_{\theta^\star}^{-1}) = \det(\mathbf{H}_{\theta^\star})^{-2} \det(\mathbf{V}_\star).$$

By Assumption 2, $\mathbf{H}_{\theta^\star}$ is a constant, deterministic matrix independent of π. Therefore, minimizing $\det(\mathbf{\Sigma}_\star)$ is strictly equivalent to minimizing $\det(\mathbf{V}_\star)$, or equivalently, minimizing the log-determinant $\log\det(\mathbf{V}_\star)$. Recall the decomposition of $\mathbf{V}_\star$ (which is a functional of π)

$$\mathbf{V}_\star(\pi) = \mathrm{var}\left(\nabla \ell_{\theta^\star}(X,Y)\right) + \mathbb{E}\left[\left(\frac{1}{\pi(X)} - 1\right) \mathbf{\Delta}_\lambda(X,Y) \mathbf{\Delta}_\lambda(X,Y)^\top\right].$$

We formulate the constrained optimization problem using a Lagrange multiplier $\gamma > 0$ for the budget constraint. The Lagrangian is

$$\mathcal{L}(\pi, \gamma) = \log\det(\mathbf{V}_\star(\pi)) + \gamma\left(\mathbb{E}[\pi(X)] - \rho\right).$$

To find the optimal sampling function π, we use the calculus of variations. Consider a small perturbation $\delta\pi(X)$. The corresponding variation in the variance matrix, $\delta\mathbf{V}_\star$, is

$$\delta\mathbf{V}_\star = \mathbb{E}\left[-\frac{\delta\pi(X)}{\pi(X)^2} \mathbf{\Delta}_\lambda(X,Y) \mathbf{\Delta}_\lambda(X,Y)^\top\right].$$

By Jacobi's formula, the variation of the log-determinant is $\delta(\log\det\mathbf{V}_\star) = \mathrm{Tr}(\mathbf{V}_\star^{-1}\delta\mathbf{V}_\star)$. Substituting $\delta\mathbf{V}_\star$ into this trace yields

$$\delta(\log\det(\mathbf{V}_\star)) = \mathbb{E}\left[-\frac{\delta\pi(X)}{\pi(X)^2} \mathrm{tr}\left(\mathbf{V}_\star^{-1} \mathbf{\Delta}_\lambda(X,Y) \mathbf{\Delta}_\lambda(X,Y)^\top\right)\right].$$

The inner term now becomes the $\mathbf{V}_\star^{-1}$-weighted squared norm of $\mathbf{\Delta}_\lambda$. We condition on X using the law of iterated expectations

$$\delta(\log\det(\mathbf{V}_\star)) = \mathbb{E}_X\left[-\frac{\delta\pi(X)}{\pi(X)^2} \mathbb{E}_{Y|X}\left\{\left\|\mathbf{V}_\star^{-1/2} \mathbf{\Delta}_\lambda(X,Y)\right\|_2^2 \,\middle|\, X\right\}\right].$$

Let $\varphi_D(X)^2 = \mathbb{E}_{Y|X}\left[\left\|\mathbf{V}_\star^{-1/2} \mathbf{\Delta}_\lambda(X,Y)\right\|_2^2 \,\middle|\, X\right]$. The variation of the full Lagrangian with respect to $\delta\pi(X)$ is:

$$\delta\mathcal{L} = \mathbb{E}_X\left[\delta\pi(X)\left(-\frac{\varphi_D(X)^2}{\pi(X)^2} + \gamma\right)\right].$$

For π_{opt} to be a stationary point, this variation must be zero for any valid perturbation $\delta\pi(X)$. This implies the term in the parentheses must be zero almost everywhere:

$$\frac{\varphi_D(X)^2}{\pi^\star(X)^2} = \gamma \implies \pi_{\mathrm{opt}}(X) = \frac{\varphi_D(X)}{\sqrt{\gamma}}.$$

To determine the constant $\sqrt{\gamma}$, we enforce the budget constraint tightly: $\mathbb{E}[\pi_{opt}(X)] = \rho$. Taking the expectation of both sides gives $\mathbb{E}[\varphi_D(X)]/\sqrt{\gamma} = \rho$, which means $\sqrt{\gamma} = \mathbb{E}[\varphi_D(X)]/\rho$.Substituting this back into the stationary condition yields the optimal rule:

$$\pi_{opt} = \rho \frac{\varphi_D(x)}{\mathbb{E}[\varphi_D(X)]}.$$

□

Proof of Theorem 2 Before starting the proof, note that we will use d/dx notation for scalar differentiation, and ∇ for multivariate. Our goal is to minimize $\det(\mathbf{\Sigma}_\star)$ with respect to $\lambda \in \mathbb{R}$. From Theorem 1, we have $\mathbf{\Sigma}_\star = \mathbf{H}_{\theta^\star}^{-1}\mathbf{V}_\star(\lambda)\mathbf{H}_{\theta^\star}^{-1}$. Using the multiplicative property of the determinant, we have

$$\det(\mathbf{\Sigma}_\star) = \det(\mathbf{H}_{\theta^\star})^{-2}\det(\mathbf{V}_\star(\lambda)).$$

Because Assumption 2 dictates that $\mathbf{H}_{\theta^\star}$ is a fixed matrix independent of λ, minimizing $\det(\mathbf{\Sigma}_\star)$ is strictly equivalent to minimizing $\det(\mathbf{V}_\star(\lambda))$. Because the logarithm is a strictly increasing function, we can equivalently minimize the log-determinant, $\log\det\mathbf{V}_\star(\lambda)$. Recall the formula for the score variance matrix

$$\mathbf{V}_\star(\lambda) = \text{var}\left(\nabla\ell_{\theta^\star}(X,Y)\right) + \mathbb{E}\left[\left(\frac{1}{\pi(X)} - 1\right)\mathbf{\Delta}_\lambda(X,Y)\mathbf{\Delta}_\lambda(X,Y)^\top\right],$$

where $\mathbf{\Delta}_\lambda(X,Y) = \nabla\ell_{\theta^\star}(X,Y) - \lambda\nabla\ell^f_{\theta^\star}(X)$. Again let $W(X) = \frac{1}{\pi(X)} - 1 \geq 0$ for notational brevity. To find the stationary point, we take the derivative of the log-determinant of λ. First, we differentiate the variance matrix with respect to λ. Because the first term, Var $(\nabla\ell_{\theta^\star})$, is independent of λ, it vanishes. We pass the derivative inside the expectation for the second term

$$\frac{d\mathbf{V}_\star(\lambda)}{d\lambda} = \mathbb{E}\left[W(X)\frac{d}{d\lambda}\left(\mathbf{\Delta}_\lambda(X,Y)\mathbf{\Delta}_\lambda(X,Y)^\top\right)\right].$$

Applying the product rule to the outer product

$$\frac{d}{d\lambda}\left(\mathbf{\Delta}_\lambda\mathbf{\Delta}_\lambda^\top\right) = \left(\frac{d\mathbf{\Delta}_\lambda}{d\lambda}\right)\mathbf{\Delta}_\lambda^\top + \mathbf{\Delta}_\lambda\left(\frac{d\mathbf{\Delta}_\lambda}{d\lambda}\right)^\top.$$

Since $\frac{d\mathbf{\Delta}_\lambda}{d\lambda} = -\nabla\ell^f_{\theta^\star}(X)$, we have

$$\frac{d\mathbf{V}_\star(\lambda)}{d\lambda} = -\mathbb{E}\left[W(X)\left(\nabla\ell^f_{\theta^\star}(X)\mathbf{\Delta}_\lambda(X,Y)^\top + \mathbf{\Delta}_\lambda(X,Y)\nabla\ell^f_{\theta^\star}(X)^\top\right)\right].$$

Now, substitute this matrix derivative back into Jacobi's formula

$$\frac{d}{d\lambda}\log(\det\mathbf{V}_\star(\lambda)) = -\mathbb{E}\left[W(X)\text{tr}\left(\mathbf{V}_\star^{-1}\left(\nabla\ell^f_{\theta^\star}\mathbf{\Delta}_\lambda^\top + \mathbf{\Delta}_\lambda(\nabla\ell^f_{\theta^\star})^\top\right)\right)\right].$$

Because the trace is a linear operator and utilizing the property that the trace of a scalar/matrix is equal to the trace of its transpose, the two terms inside the inner parenthesis are identical in trace. Thus,

$$\frac{d}{d\lambda}\log\det(\mathbf{V}_\star(\lambda)) = -2\,\mathbb{E}\left[W(X)\text{tr}\left(\mathbf{V}_\star^{-1}\mathbf{\Delta}_\lambda(\nabla\ell^f_{\theta^\star})^\top\right)\right].$$

Using the cyclic permutation property of the trace $\text{tr}(Abc^\top) = c^\top Ab$, we can rewrite the expression inside the expectation as a scalar quadratic form

$$\frac{d}{d\lambda}\log\det(\mathbf{V}_\star(\lambda)) = -2\,\mathbb{E}\left[W(X)\nabla\ell^f_{\theta^\star}(X)^\top\mathbf{V}_\star(\lambda)^{-1}\mathbf{\Delta}_\lambda(X,Y)\right].$$

To find the minimum, we set the derivative to zero. Substituting $\mathbf{\Delta}_\lambda(X,Y) = \nabla\ell_{\theta^\star}(X,Y) - \lambda\nabla\ell^f_{\theta^\star}(X)$ yields

$$\mathbb{E}\left[W(X)\nabla\ell^f_{\theta^\star}(X)^\top\mathbf{V}_\star^{-1}\Big(\nabla\ell_{\theta^\star}(X,Y) - \lambda\nabla\ell^f_{\theta^\star}(X)\Big)\right] = 0.$$

Using the linearity of expectation, and solving for λ produces the implicit identity

$$\lambda_{\text{opt}} = \frac{\mathbb{E}\left[W(X)\nabla\ell^f_{\theta^\star}(X)^\top\mathbf{V}_\star(\lambda^\star)^{-1}\nabla\ell_{\theta^\star}(X,Y)\right]}{\mathbb{E}\left[W(X)\nabla\ell^f_{\theta^\star}(X)^\top\mathbf{V}_\star(\lambda^\star)^{-1}\nabla\ell^f_{\theta^\star}(X)\right]}.$$

Finally, re-substituting $W(X) = \frac{1}{\pi(X)} - 1$ yields the final result. □

Empirical fixed-point equation for $\widehat{\lambda}_{\text{opt}}$ Applying the same argument to the empirical score variance $\widehat{\mathbf{V}}(\lambda)$ of Section 3.3, setting the derivative of $\log\det\widehat{\mathbf{V}}(\lambda)$ to zero characterizes the empirical D-optimal power-tuning parameter by the implicit fixed-point equation

$$\widehat{\lambda}_{\text{opt}} = \frac{\sum_{i=1}^{n} \nabla \ell^f_{\widetilde{\theta}_0}(X_i)^\top \widehat{\mathbf{V}}(\widehat{\lambda}_{\text{opt}})^{-1} \nabla \ell_{\widetilde{\theta}_0}(X_i, Y_i) \frac{\xi_i}{\widehat{\pi}^{(\tau)}(X_i)} \left(\frac{1}{\widehat{\pi}^{(\tau)}(X_i)} - 1\right)}{\sum_{i=1}^{n} \nabla \ell^f_{\widetilde{\theta}_0}(X_i)^\top \widehat{\mathbf{V}}(\widehat{\lambda}_{\text{opt}})^{-1} \nabla \ell^f_{\widetilde{\theta}_0}(X_i) \frac{\xi_i}{\widehat{\pi}^{(\tau)}(X_i)} \left(\frac{1}{\widehat{\pi}^{(\tau)}(X_i)} - 1\right)}. \tag{A.1}$$

The matrix $\widehat{\mathbf{V}}(\cdot)$ on the right-hand side depends on $\widehat{\lambda}_{\text{opt}}$ itself, so the equation is circular. Solving it iteratively is computationally fragile and susceptible to variance inflation in finite samples, which is why Section 3.3 minimizes the bounded objective $\log\det\{\widehat{\mathbf{V}}(\lambda) + \epsilon\mathbf{I}\}$ over $\lambda \in [0, 1]$ directly.

Derivations of A-Optimal Baselines While our framework focuses on D-optimal criteria, we also consider A-optimal active sampling and power-tuning as robust baselines. For completeness, we provide the derivations for the A-optimal parameters below.

(1) *A-Optimal Sampling ($\pi^{\text{A}}_{\text{opt}}$):* By Assumption 2, the Hessian $\mathbf{H}_{\theta^\star}$ is a constant, positive definite matrix. By Lemma 1, the asymptotic covariance is $\mathbf{\Sigma}_\star = \mathbf{H}^{-1}_{\theta^\star}\mathbf{V}_\star\mathbf{H}^{-1}_{\theta^\star}$. Using the cyclic property of the trace, our objective is to minimize

$$\text{tr}(\mathbf{\Sigma}_\star) = \text{tr}\left(\mathbf{V}_\star\mathbf{H}^{-2}_{\theta^\star}\right).$$

Using the decomposition of $\mathbf{V}_\star$ from the previous proofs and substituting this back into the trace objective, minimizing $\text{tr}(\mathbf{\Sigma}_\star)$ is equivalent to minimizing the π-dependent term

$$\min_\pi \mathbb{E}\left\{\frac{1}{\pi(X)}\text{tr}\left(\mathbf{\Delta}_\lambda(X,Y)\mathbf{\Delta}_\lambda(X,Y)^\top\mathbf{H}^{-2}_{\theta^\star}\right)\right\}.$$

Using the trace property $\text{tr}(uu^\top A) = u^\top Au = \|A^{1/2}u\|_2^2$, the inner trace simplifies strictly to the weighted Euclidean norm

$$\text{tr}\left(\mathbf{\Delta}_\lambda(X,Y)\mathbf{\Delta}_\lambda(X,Y)^\top\mathbf{H}^{-2}_{\theta^\star}\right) = \left\|\mathbf{H}^{-1}_{\theta^\star}\mathbf{\Delta}_\lambda(X,Y)\right\|_2^2.$$

We apply the law of iterated expectations to condition on the features X

$$\text{Objective} \propto \mathbb{E}_X\left[\frac{1}{\pi(X)}\mathbb{E}_{Y|X}\left\{\left\|\mathbf{H}^{-1}_{\theta^\star}\mathbf{\Delta}_\lambda(X,Y)\right\|_2^2 \,\middle|\, X\right\}\right] = \mathbb{E}\left\{\frac{\varphi_A(X)^2}{\pi(X)}\right\}.$$

Applying the Cauchy–Schwarz inequality to the budget-constrained problem

$$\mathbb{E}\left(\frac{\varphi_A(X)^2}{\pi(X)}\right)\mathbb{E}[\pi(X)] \geq \left\{\mathbb{E}\left(\frac{\varphi_A(X)}{\sqrt{\pi(X)}}\sqrt{\pi(X)}\right)\right\}^2 = \left\{\mathbb{E}(\varphi_A(X))\right\}^2.$$

The minimum is achieved if and only if $\pi(X) \propto \varphi_A(X)$. To satisfy the budget constraint tightly, we set the normalizing constant to yield

$$\pi^{\text{A}}_{\text{opt}} = \rho\frac{\varphi_A(x)}{\mathbb{E}[\varphi_A(X)]}.$$

(2) *A-Optimal Power Tuning ($\lambda^{\text{A}}_{\text{opt}}$):* To minimize $\text{tr}(\mathbf{\Sigma}_\star)$ with respect to λ, we again minimize $\text{tr}(\mathbf{V}_\star\mathbf{H}^{-2}_{\theta^\star})$. Letting $W(X) = \frac{1}{\pi(X)} - 1$ as defined previously, the objective becomes a simple quadratic in λ:

$$J(\lambda) = \mathbb{E}\left[W(X)\left(\left\|\mathbf{H}^{-1}_{\theta^\star}\nabla\ell_{\theta^\star}\right\|_2^2 - 2\lambda\left(\mathbf{H}^{-1}_{\theta^\star}\nabla\ell_{\theta^\star}\right)^\top\left(\mathbf{H}^{-1}_{\theta^\star}\nabla\ell^f_{\theta^\star}\right) + \lambda^2\left\|\mathbf{H}^{-1}_{\theta^\star}\nabla\ell^f_{\theta^\star}\right\|_2^2\right)\right].$$

Differentiating $J(\lambda)$ with respect to λ and setting to zero:

$$-\mathbb{E}\left[W(X)\left(\mathbf{H}^{-1}_{\theta^\star}\nabla\ell_{\theta^\star}\right)^\top\left(\mathbf{H}^{-1}_{\theta^\star}\nabla\ell^f_{\theta^\star}\right)\right] + \lambda\mathbb{E}\left[W(X)\left\|\mathbf{H}^{-1}_{\theta^\star}\nabla\ell^f_{\theta^\star}\right\|_2^2\right] = 0.$$

Solving for λ yields the optimal value

$$\lambda_{\text{opt}}^{\text{A}} = \frac{\mathbb{E}\left[W(X)\big(\mathbf{H}_{\theta^\star}^{-1}\nabla\ell_{\theta^\star}(X,Y)\big)^\top\big(\mathbf{H}_{\theta^\star}^{-1}\nabla\ell_{\theta^\star}^f(X)\big)\right]}{\mathbb{E}\left[W(X)\left\|\mathbf{H}_{\theta^\star}^{-1}\nabla\ell_{\theta^\star}^f(X)\right\|_2^2\right]}.$$

The second derivative is $2\,\mathbb{E}\left[W(X)\left\|\mathbf{H}_{\theta^\star}^{-1}\nabla\ell_{\theta^\star}^f(X)\right\|_2^2\right] \geq 0$, guaranteeing a global minimum. □

C Majorize-Minimize Characterization of D-Optimal Design

This section formalizes the majorize-minimize (MM) characterization of the D-optimal sampling policy summarized in Remark 3. Write $\Delta_\lambda(X,Y) = \nabla\ell_{\theta^\star}(X,Y) - \lambda\nabla\ell_{\theta^\star}^f(X)$, $M_\lambda(x) = \mathbb{E}[\Delta_\lambda\Delta_\lambda^\top \mid X = x] \succeq 0$, and $V_0 = \mathrm{Var}(\nabla\ell_{\theta^\star}(X,Y)) \succ 0$. From Appendix B, for any π, the asymptotic score variance is

$$V(\pi) = V_0 + \mathbb{E}\left[\left(\frac{1}{\pi(X)} - 1\right)M_\lambda(X)\right],\ F(\pi) = \log\det V(\pi),$$

and minimizing $\det(\Sigma_\lambda^\pi)$ is equivalent to minimizing $F(\pi)$. The feasible set at budget $\rho = \lim_{n_b,n\to\infty} n_b/n \in (0,1)$ is the convex set $\Pi_\rho = \{\pi : \mathcal{X} \to (0,1] : \mathbb{E}[\pi(X)] \leq \rho\}$. For any $A \succ 0$, define the A-whitened nuisance as $u_A(x) = \sqrt{\mathrm{tr}(A^{-1}M_\lambda(x))} = \sqrt{\mathbb{E}[||\Delta_\lambda||_{A^{-1}}^2 \mid X = x]}$, and let $\mathcal{T}(A) = \pi_A$ be the water-filled policy

$$\pi_A(x) = \min\{\eta_A u_A(x), 1\},$$

with $\eta_A > 0$ satisfying $\mathbb{E}[\min\{\eta_A u_A(X), 1\}] = \rho$.

Assumption 4 (M-M Assumptions)**.** *Let V_0, M_λ, and ρ be such that:*

- $V_0 \succ 0$*;*
- $\mathbb{E}[\mathrm{tr}(M_\lambda(X))] < \infty$*, $M_\lambda(X) \neq 0$ a.s.; and*
- $\rho \in (0,1)$.

Proposition C.1 (Majorize-Minimize Characterization of D-Optimal Design)**.** *Fix λ and suppose Assumption 4 holds. Then*

(i) ***Tangent majorization.*** *For every $A \succ 0$ and every $\pi \in \Pi_\rho$, $F(\pi) \leq Q(\pi \mid A) = \log\det A - p + \mathrm{tr}(A^{-1}V(\pi))$, with equality if and only if $V(\pi) = A$.*

(ii) ***The surrogate is minimized in closed form by water-filling.***

$$Q(\pi \mid A) = c(A) + \mathbb{E}\left[\frac{u_A(X)^2}{\pi(X)}\right],$$

where $c(A)$ does not depend on π. The level η_A exists, and $\pi_A = \mathcal{T}(A)$ is the unique minimizer of $Q(\cdot \mid A)$ over Π_ρ.

(iii) ***Monotone descent.*** *For any $\pi^{(0)} \in \Pi_\rho$, alternate $A_k = V(\pi^{(k)})$ and $\pi^{(k+1)} = \mathcal{T}(A_k)$, that is, $\pi^{(k+1)}(x) = \min\{\eta_{A_k}u_{A_k}(x), 1\}$. Then $F(\pi^{(k+1)}) \leq F(\pi^{(k)})$ for all k, with equality if and only if $\pi^{(k+1)} = \pi^{(k)}$ a.e. Consequently, $F(\pi^{(k)})$ decreases to a limit $F_\infty \geq \log\det V_0 > -\infty$.*

(iv) ***Fixed points are the Theorem 1 policies, and are stationary.*** *A policy $\pi \in \Pi_\rho$ satisfies $\pi = \mathcal{T}(V(\pi))$ if and only if*

$$\pi(x) = \min\left\{\eta\sqrt{\mathbb{E}[||\Delta_\lambda||_{V(\pi)^{-1}}^2 \mid X = x]}, 1\right\},\quad \mathbb{E}[\pi(X)] = \rho,$$

that is, the capped form of (3.4)*. Every such π is a first-order stationary point of F over Π_ρ.*

(v) ***Guarantee for the one-step mixed policy.*** *Let $\pi_{\mathrm{unif}} \equiv \rho$, let $A_0 = V(\pi_{\mathrm{unif}})$, and for $\tau \in [0,1)$, let $\pi^{(\tau)} = (1-\tau)\mathcal{T}(A_0) + \tau\rho$ be the one-step τ-mixed policy of the form* (3.5)*. Then $\pi^{(\tau)} \in \Pi_\rho$ and $F(\pi^{(\tau)}) \le F(\pi_{\mathrm{unif}})$. The one-step mixed policy never increases the asymptotic covariance determinant relative to uniform sampling, for every mixing weight τ and with the budget cap active.*

Remark 5. *Part (v) is stated for the one-step update $\mathcal{T}(A_0)$ from uniform sampling. In our experiments, we run $K = 10$ MM iterations before mixing. Starting from $\pi^{(0)} = \pi_{\mathrm{unif}}$, part (iii) gives $F(\pi^{(K)}) \le F(\mathcal{T}(A_0)) \le F(\pi_{\mathrm{unif}})$, so additional iterations cannot increase F of the unmixed policy. The proposition does not, however, cover the mixed policy $(1-\tau)\pi^{(K)} + \tau\rho$ for $K > 1$.*

Proof of Proposition C.1 We decompose this proof into the following steps.

(i) Tangent majorization. Since $1/\pi \ge 1$ and $M_\lambda \succeq 0$, $V(\pi) \succeq V_0 \succ 0$, so $A^{-1/2}V(\pi)A^{-1/2}$ has strictly positive eigenvalues $\mu_1, \dots, \mu_p$. Then,

$$\begin{aligned} Q(\pi \mid A) - F(\pi) &= \operatorname{tr}(A^{-1}V(\pi)) - p - \log\det(A^{-1}V(\pi)) \\ &= \sum_{j=1}^{p} (\mu_j - 1 - \log\mu_j) \\ &\ge 0. \end{aligned}$$

since $\log t \le t - 1$ for $t > 0$. Equality forces $\mu_j = 1$, $\forall j$, that is, $A^{-1/2}V(\pi)A^{-1/2} = I$, or $V(\pi) = A$.

(ii) Exact minimization of the surrogate. Since $\operatorname{tr}(A^{-1}M_\lambda(x)) = u_A(x)^2$,

$$\operatorname{tr}(A^{-1}V(\pi)) = \operatorname{tr}(A^{-1}V_0) - \mathbb{E}[u_A(X)^2] + \mathbb{E}\left[\frac{u_A(X)^2}{\pi(X)}\right],$$

which gives the stated form of Q with $c(A)$ free of π. Minimizing $Q(\cdot \mid A)$ over Π_ρ therefore reduces to the scalar problem of minimizing $\mathbb{E}[u_A(X)^2/\pi(X)]$ over $\pi \in \Pi_\rho$.

Now, we first prove the existence of η and follow up with the optimality. Let $g(\eta) = \mathbb{E}[\min\{\eta u_A(X), 1\}]$. It is continuous and non-decreasing, with $g(0) = 0$, so by monotone convergence, $g(\eta) \to P(u_A(X) > 0) = 1$ as $\eta \to \infty$, since $u_A > 0$ a.s. by Assumption 4 and g is strictly increasing whenever $g < 1$. As $\rho < 1$, there is a unique η_A with $g(\eta_A) = \rho$ and $\pi_A \in \Pi_\rho$ with $\mathbb{E}[\pi_A(X)] = \rho$.

For optimality, start by writing $u = u_A$, $\eta = \eta_A$, $\pi^\star = \pi_A$, and let $\pi \in \Pi_\rho$ be arbitrary. For fixed x, the map $t \mapsto u(x)^2/t$ is convex on $(0,\infty)$, so it lies above its tangent at $t = \pi^\star(x)$:

$$\frac{u(x)^2}{\pi(x)} \ge \frac{u(x)^2}{\pi^\star(x)} - \frac{u(x)^2}{\pi^\star(x)^2}(\pi(x) - \pi^\star(x)).$$

Call this the tangent bound. Now, we replace the coefficient $u(x)^2/\pi^\star(x)^2$ by the constraint $1/\eta^2$, splitting on the cap:

- Unsaturated, $\pi^\star(x) = \eta u(x) < 1$. Here, $u(x)^2/\pi^\star(x)^2 = 1/\eta^2$ exactly.
- Saturated, $\pi^\star(x) = 1 \le \eta u(x)$. Here,

  $$\frac{u(x)^2}{\pi^\star(x)^2} = u(x)^2 \ge \frac{1}{\eta^2},$$

  while $\pi(x) - \pi^\star(x) = \pi(x) - 1 \le 0$, so

  $$\left(\frac{u(x)^2}{\pi^\star(x)^2} - \frac{1}{\eta^2}\right)(\pi(x) - \pi^\star(x)) \le 0.$$

In both cases, the right-hand side of the tangent bound is at least $u(x)^2/\pi^\star(x) - \eta^{-2}(\pi(x) - \pi^\star(x))$. Taking expectations,

$$\mathbb{E}\left[\frac{u(X)^2}{\pi(X)}\right] \ge \mathbb{E}\left[\frac{u(X)^2}{\pi^\star(X)}\right] - \frac{1}{\eta^2}(\mathbb{E}[\pi(X)] - \rho) \ge \mathbb{E}\left[\frac{u(X)^2}{\pi^\star(X)}\right],$$

since $\mathbb{E}[\pi(X)] \le \rho$. For uniqueness, $t \mapsto u(x)^2/t$ is strictly convex whenever $u(x) > 0$, so the tangent bound is strict on $\{\pi \ne \pi^\star\}$, which has measure zero at any minimizer.

(iii) Monotone descent. Let $A_k = V(\pi^{(k)})$ and $\pi^{(k+1)} = \mathcal{T}(A_k)$. Then,

$$F(\pi^{(k+1)}) \le Q(\pi^{(k+1)} \mid A_k) \le Q(\pi^{(k)} \mid A_k) = F(\pi^{(k)}).$$

The first inequality is part (i); the second holds since $\pi^{(k+1)}$ minimizes $Q(\cdot \mid A_k)$ by (ii); the final equality is the equality case of part (i), which is usable since $A_k = V(\pi^{(k)})$ makes the majorizer tight at $\pi^{(k)}$.

For the lower bound, $V(\pi) \succeq V_0 \succ 0$ implies $\det V(\pi) \ge \det V_0$, so $F \ge \log\det V_0$ and the non-increasing sequence $F(\pi^{(k)})$ converges. If $F(\pi^{(k+1)}) = F(\pi^{(k)})$, the middle inequality is an equality, and uniqueness in part (ii) gives $\pi^{(k+1)} = \pi^{(k)}$ a.e.

(iv) Fixed points and stationarity. The equation in item (iv) of the proposition is the definition of $\pi = \mathcal{T}(V(\pi))$, using

$$\operatorname{tr}(V(\pi)^{-1} M_\lambda(x)) = \mathbb{E}[||\Delta_\lambda||^2_{V(\pi)^{-1}} \mid X = x].$$

When the cap is inactive, it reduces to $\pi(x) = \eta\sqrt{\mathbb{E}[\|\Delta_\lambda\|^2_{V(\pi)^{-1}} \mid X = x]}$ with η fixed by the budget, which is Eq. (3.4).

For stationarity, note that $Q(\cdot \mid A)$ is tangent to F at any π with $V(\pi) = A$: for a feasible direction $h = \pi' - \pi$ for $\pi' \in \Pi_\rho$,

$$DF(\pi)[h] = \operatorname{tr}(V(\pi)^{-1} DV(\pi)[h]) = DQ(\pi \mid V(\pi))[h],$$
$$DV(\pi)[h] = -\mathbb{E}\left[\frac{h(X)}{\pi(X)^2} M_\lambda(X)\right].$$

Both derivatives are finite, since at a fixed point

$$\frac{u^2_{V(\pi)}}{\pi^2} \le \max\{\eta^{-2}, u^2_{V(\pi)}\},$$

which is integrable under Assumption 4. If π is a fixed point then it minimizes the convex functional $Q(\cdot \mid V(\pi))$ over the convex set Π_ρ, so

$$DQ(\pi \mid V(\pi))[\pi' - \pi] \ge 0, \quad \forall \pi' \in \Pi_\rho,$$

and hence $DF(\pi)[\pi' - \pi] \ge 0$: no feasible direction decreases F to first order. On the other hand, a policy satisfying this variational inequality with the active budget constraint satisfies the equation in item (iv) of the proposition, by the same argument as in part (ii).

(v) The one-step mixed policy. Π_ρ is convex and $\pi^{(\tau)} = (1-\tau)\mathcal{T}(A_0) + \tau\pi_{\text{unif}}$ is a convex combination of its elements, so $\pi^{(\tau)} \in \Pi_\rho$. By item (ii) of this proposition, $\pi \mapsto Q(\pi \mid A_0) = c(A_0) + \mathbb{E}[u^2_{A_0}/\pi]$ is convex, so

$$Q(\pi^{(\tau)} \mid A_0) \le (1-\tau)Q(\mathcal{T}(A_0) \mid A_0) + \tau Q(\pi_{\text{unif}} \mid A_0) \le Q(\pi_{\text{unif}} \mid A_0) = F(\pi_{\text{unif}}),$$

where the second inequality uses the minimality of $\mathcal{T}(A_0)$ and the final equality is the equality case of part (i) with $A_0 = V(\pi_{\text{unif}})$. Combining with part (i) applied to $\pi^{(\tau)}$ gives

$$F(\pi^{(\tau)}) \le Q(\pi^{(\tau)} \mid A_0) \le F(\pi_{\text{unif}}).$$

□

D Proofs for Asymptotic Behavior

Proof of Lemma 1 Recall the empirical optimal active loss function $L^\pi_\lambda(\theta)$ is defined as

$$L^\pi_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \left\{\lambda \ell^f_{\theta,i} + \left(\ell_{\theta,i} - \lambda \ell^f_{\theta,i}\right)\frac{\xi_i}{\pi(X_i)}\right\}.$$

To show the unbiasedness of $L^{\pi}_{\lambda}(\theta)$, we evaluate the expectation of $L^{\pi}_{\lambda}(\theta)$ for any tuning parameter λ, and sampling policy π. The randomness comes from both the data $\mathcal{D}$ and the active sampling mechanism $\xi_i \overset{\text{i.i.d}}{\sim} \text{Bern}(\pi(X_i)),\ i \in [n]$.

By the law of total expectation, conditioning on the full data $\mathcal{D}$, and noting that the sampling indicator ξ_i only depends on X_i, we have $\mathbb{E}\left\{\xi_i \mid X_i, Y_i\right\} = \mathbb{E}\left\{\xi_i \mid X_i\right\} = \pi(X_i)$. Therefore,

$$\begin{aligned}
\mathbb{E}\left\{L^{\pi}_{\lambda}(\theta)\right\} &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\xi}\left\{L^{\pi}_{\lambda}(\theta) \mid \mathcal{D}\right\}\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\lambda\ell^{f}_{\theta,i} + (\ell_{\theta,i} - \lambda\ell^{f}_{\theta,i})\frac{\mathbb{E}_{\xi}[\xi_i \mid X_i]}{\pi(X_i)}\right\}\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\lambda\ell^{f}_{\theta,i} + (\ell_{\theta,i} - \lambda\ell^{f}_{\theta,i})\frac{\pi(X_i)}{\pi(X_i)}\right\}\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\lambda\ell^{f}_{\theta,i} + \ell_{\theta,i} - \lambda\ell^{f}_{\theta,i}\right\}\right] \\
&= \mathbb{E}_{\mathcal{D}}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell_{\theta}(X_i, Y_i)\right\} \\
&= \mathbb{E}\left\{\ell_{\theta}(X, Y)\right\}.
\end{aligned}$$

This confirms that the active empirical risk $L^{\pi}_{\lambda}(\theta)$ is an unbiased estimator of the true population risk for any $\theta \in \Theta$, establishing the foundation for the consistency and asymptotic normality of the resulting M-estimator $\widehat{\theta}^{\pi}_{\lambda}$. □

Proof of Theorem 3 (M-estimation) The proof follows from that of the classical proof of asymptotic normality for M-estimation in Vaart (1998). Two very similar proofs are also given for the original active statistical inference estimator (Zrnic and Candès, 2024) and the prediction-powered estimator (Angelopoulos et al., 2023b).

Motivated by Zrnic and Candès (2024), for any $\eta \in \mathbb{R}^{+}$, we define

$$\xi^{\eta}_i = 1\{\pi_{\eta}(X_i) \le \pi_{\hat{\eta}}(X_i)\}\xi_i(1 - \xi^{\leq}_i) + 1\{\pi_{\eta}(X_i) > \pi_{\hat{\eta}}(X_i)\}(\xi_i + (1 - \xi_i)\xi^{>}_i), \tag{A.1}$$

where

$$\xi^{\leq}_i \sim \text{Bern}\left(\frac{\pi_{\hat{\eta}}(X_i) - \pi_{\eta}(X_i)}{\pi_{\hat{\eta}}(X_i)}\right),\ \ \xi^{>}_i \sim \text{Bern}\left(\frac{\pi_{\eta}(X_i) - \pi_{\hat{\eta}}(X_i)}{1 - \pi_{\hat{\eta}}(X_i)}\right),$$

are drawn independently of ξ_i. Then $\xi^{\eta^\star}_i$ is coupled with ξ_i while ensuring $\xi^{\eta^\star}_i \sim \text{Bern}(\pi_{\eta^\star}(X_i))$. Define

$$L^{\eta,\lambda}_{\theta,i} = \lambda\ell^{f}_{\theta,i} + (\ell_{\theta,i} - \lambda\ell^{f}_{\theta,i})\frac{\xi_i}{\pi(X_i)},$$

and given some function g, let

$$\begin{aligned}
\mathbb{G}_n[g(L^{\eta,\lambda}_{\theta})] &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(g(L^{\eta,\lambda}_{\theta,i}) - \mathbb{E}[g(L^{\eta,\lambda}_{\theta,i})]\right); \\
\mathbb{E}_n[g(L^{\eta,\lambda}_{\theta})] &:= \frac{1}{n}\sum_{i=1}^{n} g(L^{\eta,\lambda}_{\theta,i}).
\end{aligned}$$

Denote $\mathbb{G}_n[\nabla g(L^{\eta,\lambda}_{\theta})]$, $\mathbb{E}_n[\nabla g(L^{\eta,\lambda}_{\theta})]$ be the first gradients of $\mathbb{G}_n[g(L^{\eta,\lambda}_{\theta})]$ and $\mathbb{E}_n[g(L^{\eta,\lambda}_{\theta})]$, respectively. First, take a Taylor expansion of $L^{\eta^\star,\lambda^\star}_{\theta^\star,i}$ around $\theta^\star$ with perturbation $h_n/\sqrt{n}$ for some $h_n = O_p(1)$

$$L^{\eta^\star,\lambda^\star}_{\theta^\star + h_n/\sqrt{n},i} = L^{\eta^\star,\lambda^\star}_{\theta^\star,i} + \left(\frac{h_n}{\sqrt{n}}\right)^{\top}\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star,i} + R_{n,i},$$

so that

$$\sqrt{n}(L^{\eta^\star,\lambda^\star}_{\theta^\star + h_n/\sqrt{n},i} - L^{\eta^\star,\lambda^\star}_{\theta^\star,i}) - h_n^{\top}\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star,i} = \sqrt{n}R_{n,i}.$$

Now apply $\mathbb{G}_n$ to both sides:

$$\mathbb{G}_n[\sqrt{n}(L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n},i} - L^{\eta^\star,\lambda^\star}_{\theta^\star,i}) - h_n^\top \nabla L^{\eta^\star,\lambda^\star}_{\theta^\star,i}] = \mathbb{G}_n[\sqrt{n}R_{n,i}],$$

where

$$\begin{aligned}\mathbb{G}_n[\sqrt{n}R_{n,i}] &= \frac{1}{\sqrt{n}}\sum_{i=1}^n(\sqrt{n}R_{n,i} - \mathbb{E}[\sqrt{n}R_{n,i}]) \\ &= \sum_{i=1}^n (R_{n,i} - \mathbb{E}[R_{n,i}])\,.\end{aligned}$$

Since $h_n = O_p(1)$, the perturbation shrinks at a rate of $1/\sqrt{n}$ and since the loss is locally Lipschitz, $R_{n,i} = o_p(n^{-1/2})$ holds uniformly. So,

$$\mathbb{G}_n[\sqrt{n}(L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n},i} - L^{\eta^\star,\lambda^\star}_{\theta^\star,i}) - h_n^\top \nabla L^{\eta^\star,\lambda^\star}_{\theta^\star,i}] = o_p(1),$$

which is equivalent to

$$\sqrt{n}\mathbb{G}_n[L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] - h_n^\top \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] = o_p(1).$$

Recall $\mathbb{G}_n[g] = \sqrt{n}(\mathbb{E}_n[g] - \mathbb{E}[g])$, so apply to the first term in the above:

$$\begin{aligned}\sqrt{n}\mathbb{G}_n[L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] &= n\mathbb{E}_n[L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] - n\mathbb{E}[L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] \\ &= n\mathbb{E}_n[L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] - n(L(\theta^\star + h_n/\sqrt{n}) - L(\theta^\star)).\end{aligned}$$

Combining the terms now yields

$$n\mathbb{E}_n[L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] = n(L(\theta^\star + h_n/\sqrt{n}) - L(\theta^\star)) + h_n^\top \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] + o_p(1).$$

Next, take a Taylor expansion of $L(\theta)$ around $\theta^\star$ with step size $h_n/\sqrt{n}$:

$$L(\theta^\star + h_n/\sqrt{n}) - L(\theta^\star) = \frac{1}{2n}h_n^\top H_{\theta^\star} h_n + o(n^{-1})$$

so

$$n(L(\theta^\star + h_n/\sqrt{n}) - L(\theta^\star)) = \frac{1}{2}h_n^\top H_{\theta^\star} h_n + o(1)$$

and thus

$$n\mathbb{E}_n[L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] = \frac{1}{2}h_n^\top H_{\theta^\star} h_n + h_n^\top \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] + o_p(1).$$

Now we need to deal with the data dependent terms $\hat{\eta}$ and $\hat{\lambda}$. First, since $\mathbb{P}(\hat{\eta} \neq \eta^\star) \to 0$,

$$n\mathbb{E}_n[L^{\widehat{\eta},\widehat{\lambda}}_{\theta^\star+h_n/\sqrt{n}} - L^{\widehat{\eta},\widehat{\lambda}}_{\theta^\star}] = n\mathbb{E}_n[L^{\eta^\star,\widehat{\lambda}}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\widehat{\lambda}}_{\theta^\star}] + o_p(1).$$

Second, decompose $L^{\eta^\star,\hat{\lambda}}_{\theta,i}$ as

$$L^{\eta^\star,\hat{\lambda}}_{\theta,i} = L^{\eta^\star,\lambda^\star}_{\theta,i} + (\hat{\lambda} - \lambda^\star)\ell^f_{\theta,i}\left(1 - \frac{\xi_i^{\eta^\star}}{\pi_{\eta^\star}(X_i)}\right).$$

With $\Delta_{\theta^\star}\ell_i^f = \ell^f_{\theta^\star+h_n/\sqrt{n},i} - \ell^f_{\theta^\star,i}$, we have

$$n\mathbb{E}_n[L^{\eta^\star,\widehat{\lambda}}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\widehat{\lambda}}_{\theta^\star}] = n\mathbb{E}_n[L^{\eta^\star,\lambda^\star}_{\theta^\star+h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] + (\widehat{\lambda} - \lambda^\star)\sum_{i=1}^n \Delta_{\theta^\star}\ell_i^f\left(1 - \frac{\xi_i^{\eta^\star}}{\pi_{\eta^\star}(X_i)}\right).$$

A first order Taylor expansion of $\Delta_{\theta^\star}\ell_i^f$ yields

$$\Delta_{\theta^\star}\ell_i^f = \left(\frac{h_n}{\sqrt{n}}\right)^\top \nabla\ell^f_{\theta^\star,i} + o_p(n^{-1/2}),$$

and thus

$$\sum_{i=1}^{n} \Delta_{\theta^\star} \ell_i^f \left(1 - \frac{\xi_i^{\eta^\star}}{\pi_{\eta^\star}(X_i)}\right) = h_n^\top \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \ell_{\theta^\star,i}^f \left(1 - \frac{\xi_i^{\eta^\star}}{\pi_{\eta^\star}(X_i)}\right) \right\} + o_p(1).$$

Note that $\mathbb{E}[\xi_i^{\eta^\star}/\pi_{\eta^\star}(X_i)] = 1$, the bracketed expression above has zero mean and, by the CLT, has a limit distribution, and thus is $O_p(1)$. By $h_n = O_p(1)$ and $(\widehat{\lambda} - \lambda^\star) = o_p(1)$,

$$\begin{aligned} n\mathbb{E}_n[L^{\eta^\star,\widehat{\lambda}}_{\theta^\star + h_n/\sqrt{n}} - L^{\eta^\star,\widehat{\lambda}}_{\theta^\star}] &= n\mathbb{E}_n[L^{\eta^\star,\lambda^\star}_{\theta^\star + h_n/\sqrt{n}} - L^{\eta^\star,\lambda^\star}_{\theta^\star}] + o_p(1) \\ &= \frac{1}{2} h_n^\top \mathbf{H}_{\theta^\star} h_n + h_n^\top \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] + o_p(1). \end{aligned}$$

Now let h_n take two distinct values. First,

$$h_n = \widehat{h}_n := \sqrt{n}(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}} - \theta^\star),$$

and second,

$$h_n = \widetilde{h}_n := -\mathbf{H}_{\theta^\star}^{-1} \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}].$$

By Lemma A.1, we have $\widehat{h}_n = O_p(1)$, and thus

$$n\mathbb{E}_n[L^{\widehat{\eta},\widehat{\lambda}}_{\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}}} - L^{\widehat{\eta},\widehat{\lambda}}_{\theta^\star}] = \frac{1}{2} \widehat{h}_n^\top \mathbf{H}_{\theta^\star} \widehat{h}_n + \widehat{h}_n^\top \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] + o_p(1),$$

and

$$n\mathbb{E}_n[L^{\widehat{\eta},\widehat{\lambda}}_{\theta^\star + \widetilde{h}_n/\sqrt{n}} - L^{\widehat{\eta},\widehat{\lambda}}_{\theta^\star}] = \frac{1}{2} \widetilde{h}_n^\top \mathbf{H}_{\theta^\star} \widetilde{h}_n + \widetilde{h}_n^\top \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] + o_p(1).$$

By definition, $\widehat{\theta}^{\widehat{\eta},\widehat{\lambda}}$ minimizes the empirical loss and so

$$n\mathbb{E}_n[L^{\widehat{\eta},\widehat{\lambda}}_{\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}}} - L^{\widehat{\eta},\widehat{\lambda}}_{\theta^\star}] \leq n\mathbb{E}_n[L^{\widehat{\eta},\widehat{\lambda}}_{\theta^\star + \widetilde{h}_n/\sqrt{n}} - L^{\widehat{\eta},\widehat{\lambda}}_{\theta^\star}].$$

And thus,

$$\frac{1}{2} \widehat{h}_n^\top \mathbf{H}_{\theta^\star} \widehat{h}_n + \widehat{h}_n^\top \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] \leq \frac{1}{2} \widetilde{h}_n^\top \mathbf{H}_{\theta^\star} \widetilde{h}_n + \widetilde{h}_n^\top \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] + o_p(1).$$

Next, take $\widetilde{h}_n = -\mathbf{H}_{\theta^\star}^{-1} \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}]$ and multiply both sides by $-\mathbf{H}_{\theta^\star}$ to get

$$\mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] = -\mathbf{H}_{\theta^\star} \widetilde{h}_n.$$

Now substitute this into our expression

$$\begin{aligned} \frac{1}{2} \widehat{h}_n^\top \mathbf{H}_{\theta^\star} \widehat{h}_n - \widehat{h}_n^\top \mathbf{H}_{\theta^\star} \widetilde{h}_n &\leq \frac{1}{2} \widetilde{h}_n^\top \mathbf{H}_{\theta^\star} \widetilde{h}_n - \widetilde{h}_n^\top \mathbf{H}_{\theta^\star} \widetilde{h}_n + o_p(1) \\ &= -\frac{1}{2} \widetilde{h}_n^\top \mathbf{H}_{\theta^\star} \widetilde{h}_n + o_p(1), \end{aligned}$$

and then

$$\frac{1}{2} \left(\widehat{h}_n - \widetilde{h}_n\right)^\top \mathbf{H}_{\theta^\star} \left(\widehat{h}_n - \widetilde{h}_n\right) \leq o_p(1).$$

By Assumption 2, $\mathbf{H}_{\theta^\star}$ is strictly positive semi-definite, so the above is strictly non-negative, and thus $\widehat{h}_n - \widetilde{h}_n \xrightarrow{p} 0$ such that

$$\sqrt{n}\left(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}} - \theta^\star\right) = -\mathbf{H}_{\theta^\star}^{-1} \mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] + o_p(1).$$

By CLT, $\mathbb{G}_n[\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}] \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\star)$ where

$$\mathbf{V}_\star = \mathrm{var}\left(\nabla L^{\eta^\star,\lambda^\star}_{\theta^\star}\right).$$

By Slutsky's theorem, we have

$$\sqrt{n}\left(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}} - \theta^\star\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{H}_{\theta^\star}^{-1} \mathbf{V}_\star \mathbf{H}_{\theta^\star}^{-1}\right).$$

□

Proof of Theorem 4 From Theorem 3, we have the joint asymptotic normality of the data-dependent estimator

$$\sqrt{n}\left(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda}} - \theta^\star\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma}_\star).$$

By the marginalization property of the multivariate normal distribution, the j-th coordinate satisfies:

$$\sqrt{n}\left(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda},j} - \theta^\star_j\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma}_{\star,jj}).$$

We are given that $\widehat{\mathbf{\Sigma}}$ is a consistent estimator of the asymptotic covariance matrix, meaning $\widehat{\mathbf{\Sigma}} \xrightarrow{p} \mathbf{\Sigma}_\star$. By the Continuous Mapping Theorem, $\sqrt{\widehat{\mathbf{\Sigma}}_{jj}} \xrightarrow{p} \sqrt{\mathbf{\Sigma}_{\star,jj}}$. We apply Slutsky's Theorem to form the studentized statistic:

$$Z_n = \frac{\sqrt{n}\left(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda},j} - \theta^\star_j\right)}{\sqrt{\widehat{\mathbf{\Sigma}}_{jj}}} \xrightarrow{d} \mathcal{N}(0,1).$$

By the definition of convergence in distribution to a standard normal random variable, the probability that Z_n falls within the critical values $\pm z_{1-\alpha/2}$ converges to $1-\alpha$:

$$\lim_{n\to\infty} \mathbb{P}\left(-z_{1-\alpha/2} \le \frac{\sqrt{n}\left(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda},j} - \theta^\star_j\right)}{\sqrt{\widehat{\mathbf{\Sigma}}_{jj}}} \le z_{1-\alpha/2}\right) = 1-\alpha.$$

Rearranging the inequality inside the probability yields the desired confidence interval coverage:

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda},j} - z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathbf{\Sigma}}_{jj}}{n}} \le \theta^\star_j \le \widehat{\theta}^{\widehat{\pi}}_{\widehat{\lambda},j} + z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathbf{\Sigma}}_{jj}}{n}}\right) = 1-\alpha.$$

This confirms that $\lim_{n\to\infty} \mathbb{P}\left\{\theta^\star_j \in \mathcal{C}_\alpha(\theta^\star_j)\right\} = 1-\alpha$, completing the proof. □

E Comparing Optimality Criteria

Let $\mu_1, \ldots, \mu_p > 0$ denote the eigenvalues of the asymptotic covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$. Geometrically, $\mathbf{\Sigma}$ defines the shape of the $(1-\alpha)$-level simultaneous confidence ellipsoid for the target parameter $\theta^\star$. The semi-axis lengths of this ellipsoid are proportional to $\sqrt{\mu_j}$. The standard optimality criteria optimize different summaries of these eigenvalues:

- **A-optimality** minimizes the trace, $\mathrm{tr}(\mathbf{\Sigma}) = \sum_{j=1}^p \mu_j$. It targets the average marginal variance, effectively minimizing the sum of the squared lengths of the bounding box enclosing the confidence ellipsoid.
- **D-optimality** minimizes the determinant, $\det(\mathbf{\Sigma}) = \prod_{j=1}^p \mu_j$. It directly minimizes the volume of the joint confidence ellipsoid.
- **E-optimality** minimizes the maximum eigenvalue, $\lambda_{\max}(\mathbf{\Sigma}) = \max_j \mu_j$. It minimizes the worst-case variance in any single parameter direction.

The relationship among the three optimality criteria is inherently bounded by the generalized mean inequality

$$\underbrace{\left(\prod_{j=1}^p \mu_j\right)^{1/p}}_{\text{Geometric Mean (D-opt)}} \le \underbrace{\frac{1}{p}\sum_{j=1}^p \mu_j}_{\text{Arithmetic Mean (A-opt)}} \le \underbrace{\max_j \mu_j}_{\text{Maximum (E-opt)}}, \tag{A.1}$$

where the equalities hold if and only if $\mu_1 = \cdots = \mu_p$, corresponding to a spherical confidence region. Equation (A.1) illustrates that while A-optimality controls the arithmetic average of the marginal variances, D-optimality systematically bounds this average from below by minimizing the geometric mean, thereby regularizing the overall covariance structure.

We provide a small example for illustration. Let $p = 2$ with eigenvalues $\mu_1 = 9$ and $\mu_2 = 1$, representing a highly elongated confidence ellipse with major axis 3 and minor axis 1. The effective "average variance" scores (which we wish to minimize) under the three criteria are

- **E** (Maximum): $\max(9, 1) = 9$.
- **A** (Arithmetic Mean): $\frac{9+1}{2} = 5$.
- **D** (Geometric Mean): $(9 \times 1)^{1/2} = 3$.

As visualized in Figure A1, A-optimality (score 5) masks the severe spectral imbalance ($\mu_1/\mu_2 = 9$). It can aggressively reduce variance along one axis while leaving the other heavily inflated, yielding an anisotropic ellipsoid with a large actual volume. E-optimality (score 9) is governed entirely by the worst-case direction, neglecting potential information gains in the remaining dimensions. D-optimality (score 3), by capturing the joint spread through the geometric mean, directly reflects the true volume of the uncertainty region. Geometrically, the D-equivalent circle (red) tightly hugs the core volume of the ellipse. In contrast, the A-equivalent circle and E-equivalent circle produce loose overestimations of the effective uncertainty. Since our inferential goal is the joint estimation of the full parameter vector $\theta^\star$, and the volume of the simultaneous confidence ellipsoid is proportional to $\det(\mathbf{\Sigma}_\star)^{1/2}$, D-optimality emerges as the most principled and theoretically appropriate criterion.

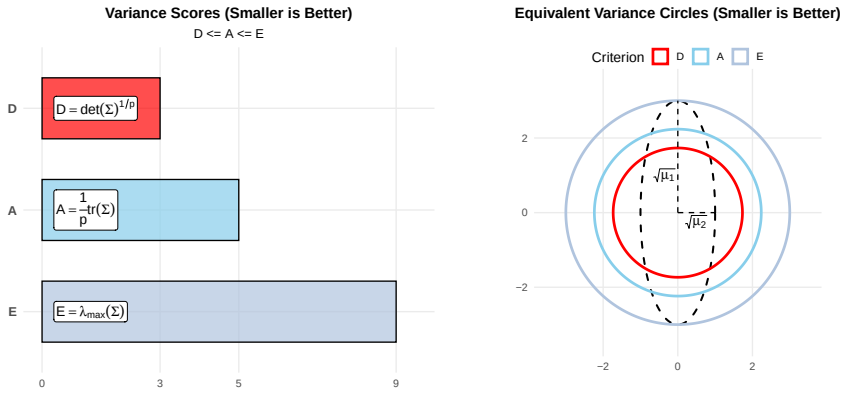


Figure A1: Comparison of different optimality criteria.

F Limitations and Broader Impacts

While the proposed `OptAPI` framework provides optimal sampling and valid theoretical guarantees under a labeling budget, it is not without limitations:

- **Distributional assumptions:** The methodology inherently assumes that the historical pilot data, the unlabeled pool, and the target population are drawn from the same underlying distribution. In settings with severe covariate shift or adversarial data contamination, the proxy model's utility may degrade. Future work could explore extending this optimal active design to domain adaptation settings to explicitly handle distribution shifts.
- **Pilot data dependence:** The finite-sample efficiency of our active sampling policy relies on the quality of the initial pilot dataset used to construct the preliminary estimate $\widehat{\theta}$ and the proxy $f(x)$. A highly restricted or non-representative pilot set may yield a poor initial covariance estimate, mildly attenuating the early efficiency gains of the active design. In our pilot-size ablation (Appendix G.7), however, a pilot of only 2% of the label budget already recovers over 99% of the efficiency of an oracle design.
- **Computational cost in high dimensions:** While the D-optimality criterion rigorously captures joint multivariate uncertainty, computing the determinant of the asymptotic covariance matrix can be computationally expensive for ultra-high-dimensional parameter spaces compared to simpler, trace-based heuristics.
- **Overhead with uninformative proxies:** While `OptAPI` is theoretically safe, guaranteeing valid confidence intervals even when the proxy is entirely misspecified, an uninformative proxy means the framework will gracefully default to the classical baseline performance, albeit having expanded the computational overhead of the two-stage tuning process.

G Additional Numerical Results and Experimental Compute Resources

This section provides additional empirical results and describes the compute resources. We first report the full logistic regression experiment, including a detailed breakdown of estimation accuracy and efficiency. We then study a Poisson regression setting to highlight behavior under severe heteroskedasticity, followed by a real-world application to census healthcare data. Finally, we present three ablation studies examining the sensitivity of `OptAPI` to its key hyperparameters and inputs: the mixing parameter τ, the initial proxy weight λ_{init}, and the pilot sample size n_l.

G.1 Compute Resources

All simulations and experiments were implemented in R version 4.4.2 (R Core Team, 2024), with Monte Carlo replications parallelized across CPU cores. All results were generated on a single Apple MacBook Air with an Apple M2 chip and 16 GB of memory; no GPU or cluster resources were used. Most experiments, each with 500 Monte Carlo replications, completed within a few minutes, and no experiment took longer than one hour.

G.2 Logistic Regression

We evaluate `OptAPI` on a binary classification task under the logistic loss, with detailed results for mean squared error (MSE), empirical coverage, and effective sample size (n_{eff}) reported in Table A1. All methods achieve empirical coverage close to the nominal 95% level. In contrast to the substantial gains observed in heteroskedastic regression settings, however, the efficiency improvements here are more modest. This is consistent with the geometry of the logistic model in the present regime: since the true conditional probabilities $p(x) = \mathbb{P}(Y = 1 \mid X = x)$ remain bounded away from 0 and 1, the variance function $p(x)(1 - p(x))$ stays relatively flat across the feature space. As a result, there are no strongly localized high-variance regions for active sampling to exploit, and the oracle sampling rule is correspondingly closer to uniform.

This setting is therefore best viewed as a stress test for the safety of `OptAPI`. When the proxy provides limited actionable heterogeneity, untuned proxy-based methods such as standard PPI and Active Inference may suffer efficiency loss under proxy misspecification. By contrast, `OptAPI` adapts its power-tuning parameter to reduce reliance on the proxy in this regime, thereby avoiding variance inflation and remaining competitive with, and slightly better than, the classical baseline.

Table A1: Performance of Logistic Regression (Coverage, MSE and n_{eff}) at selected budget levels. Values are reported as Mean (Standard deviation) across iterations.

Method	Budget $\rho = 0.1$			Budget $\rho = 0.5$			Budget $\rho = 0.9$		
	Cov (%)	MSE	n_{eff}	Cov (%)	MSE	n_{eff}	Cov (%)	MSE	n_{eff}
`OptAPI` (Ours)	**94.8 (22.2)**	**0.187 (0.231)**	**956 (85)**	**94.6 (22.6)**	**0.0356 (0.0465)**	**4769 (135)**	**95.4 (21.0)**	**0.0200 (0.0211)**	**8560 (98)**
PPI++	95.2 (21.4)	0.200 (0.253)	944 (93)	95.4 (21.0)	0.0359 (0.0457)	4732 (146)	96.2 (19.1)	0.0205 (0.0236)	8542 (80)
Classical	95.6 (20.5)	0.187 (0.253)	950 (0)	95.4 (21.0)	0.0345 (0.0446)	4750 (0)	95.8 (20.1)	0.0200 (0.0235)	8550 (0)
Confident Active Inference	94.6 (22.6)	0.223 (0.283)	948 (87)	96.6 (18.1)	0.0341 (0.0429)	4748 (150)	96.6 (18.1)	0.0202 (0.0232)	8477 (136)
Active Inference	95.6 (20.5)	0.201 (0.248)	916 (89)	96.2 (19.1)	0.0389 (0.0475)	4654 (149)	96.0 (19.6)	0.0203 (0.0225)	8451 (146)
PPI	95.8 (20.1)	0.197 (0.228)	915 (82)	94.6 (22.6)	0.0387 (0.0517)	4660 (144)	95.0 (21.8)	0.0206 (0.0241)	8512 (85)

G.3 Poisson Regression

To further demonstrate the adaptivity of the D-optimal design, we test the framework on a Poisson regression task. In this setting, the target variable $Y|X$ represents count data generated with an exponential link function, inherently exhibiting strong heteroskedasticity since the variance scales directly with the mean ($\text{var}(Y|X) = \mathbb{E}[Y|X]$).

As illustrated in Figure A2, the active sampling mechanisms in `OptAPI` aggressively exploit this structural variance. By systematically querying the labels of data points situated in high-variance regions of the feature space, `OptAPI` drastically reduces the uncertainty volume of the confidence ellipsoids, yielding substantial gains in effective sample size compared to purely uniform methods

(Classical, PPI, and PPI++). All active methods correctly maintain nominal 95% coverage, validating the theoretical soundness of the framework under exponential variance landscapes.

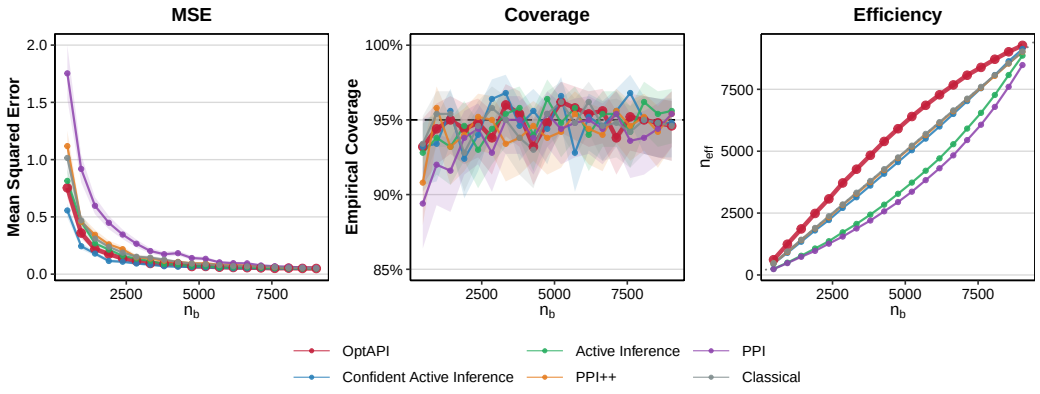


Figure A2: Poisson regression simulation results demonstrating substantial efficiency gain in heteroskedasticity settings.

G.4 Real-World Application: Census Healthcare

Dataset. We further evaluate `OptAPI` on the census healthcare benchmark of Angelopoulos et al. (2023a), constructed from 2019 American Community Survey data for California. The response Y is a binary health-insurance indicator (sample mean 0.671), and the target $\theta^\star \in \mathbb{R}^2$ is the logistic regression coefficient vector for an intercept and annual income, so that $p = 2$.

Income is heavily right-skewed, ranging from $-7{,}200$ to $1{,}403{,}700$ with a median of $27{,}800$; after standardization over the full dataset, the covariate spans approximately $[-0.73, 16.8]$. The proxy $f(X)$ is the precomputed prediction of $\mathbb{P}(Y = 1 \mid X)$ released with the benchmark (Angelopoulos et al., 2023a). It is held fixed across replications and satisfies $\mathrm{corr}\{f(X), Y\} = 0.71$.

Setup. The full dataset contains $N = 318{,}215$ individuals. Following the Full-Data Oracle methodology of Section 4.3, we define the ground truth $\theta^\star = (1.003, 1.536)^\top$ as the logistic MLE over all N rows. In each of 500 Monte Carlo replications, we draw a fresh random split into a pilot set $\mathcal{D}_L$ of $n_l = 1{,}000$ labeled points and an unlabeled pool of $n = 317{,}215$, and sweep the budget n_b/n from 5% to 95% (i.e., n_b from about 15,900 to 301,000). The pilot is used only to compute $\widetilde{\theta}_0$ and to fit the Stage I surrogate. Throughout, we set $\tau = 0.10$ in (3.5), initialize $\lambda_0 = 1$, and run ten MM iterations for the D-optimal policy. Coverage is assessed with the joint 95% Wald ellipsoid $\{\theta : (\widehat{\theta} - \theta)^\top \widehat{\mathbf{\Sigma}}^{-1} (\widehat{\theta} - \theta) \le \chi^2_{p,0.95}\}$, and n_{eff} is computed from Monte Carlo averages as $n_{\text{eff}} = n_b \, (\bar{\mathcal{V}}_{\text{classical}} / \bar{\mathcal{V}}_{\text{method}})^2$ with $\mathcal{V} = \det(\widehat{\mathbf{\Sigma}})^{1/(2p)}$, which reduces to the definition in Section 4 when $\widehat{\mathbf{\Sigma}}$ is constant across replications. By construction, Classical has $n_{\text{eff}} = n_b$, so the gain over the nominal budget coincides with the gain over Classical. In both this experiment and the pilot-size ablation (Appendix G.7), the realized labeling fraction matched the nominal budget to within 5×10^{-4}.

Surrogate construction. For linear regression and the GLMs considered here, the score factorizes as $\nabla \ell_\theta(x, y) = s_\theta(x, y)\, x$ for a scalar function s_θ, and likewise $\nabla \ell^f_\theta(x) = s^f_\theta(x)\, x$. The Stage I target therefore has rank-one structure,

$$\mathbb{E}\left\{ \left\| \nabla \ell_{\widetilde{\theta}_0}(X, Y) - \lambda_0 \nabla \ell^f_{\widetilde{\theta}_0}(X) \right\|^2_{\widetilde{\mathbf{V}}^{-1}} \,\middle|\, X = x \right\} = v(x)\, x^\top \widetilde{\mathbf{V}}^{-1} x,$$

$$v(x) = \mathbb{E}\left[\left\{ s_{\widetilde{\theta}_0}(X, Y) - \lambda_0 s^f_{\widetilde{\theta}_0}(X) \right\}^2 \,\middle|\, X = x \right],$$

so only the scalar nuisance $v(x)$ needs to be learned. We fit v on the pilot by quasi-Poisson regression with a log link, clamp its predictions to the range fitted on the pilot, floor them at a small relative

level, and set $u(x) = \{\widehat{v}(x)\, x^\top \widetilde{\mathbf{V}}^{-1} x\}^{1/2}$. The census data illustrate why a positive nuisance model matters. With a linear fit for v, extrapolation into the far right tail of income (standardized income near 10.7) produced negative values; flooring them collapsed $u(x)$ exactly at the highest-leverage points, and the resulting design was about 33% less efficient than uniform sampling. The log link guarantees $\widehat{v} > 0$, and the range clamp prevents unstable tail extrapolation.

Results. As shown in Table A2 and Figure A3, `OptAPI` attains the largest n_{eff}, equivalently the smallest confidence-ellipsoid volume, at every budget. Its gain over the nominal budget is 103.5% at $n_b/n = 5\%$, 96.2% at 10%, 76.8% at 25%, 42.3% at 50%, and 3.0% at 95%. The shrinking gain is expected, since every design converges to labeling the entire pool as $n_b \to n$. Relative to uniformly sampled PPI++, `OptAPI` improves n_{eff} by 44.2% at 5% and 20.5% at 50%. Relative to the strongest A-optimal competitor, Confident Active Inference, the improvement is 7.5% at 5%, peaks at about 9.8% at 35%, and is 8.0% at 50%. These volume differences are significant under a paired bootstrap at every budget: `OptAPI` yields the smaller volume in 92–100% of replications against every competitor. In terms of accuracy, `OptAPI` reduces MSE relative to Classical by 70.6% at 5% and 82.6% at 50%, and relative to PPI++ by 63.4% and 76.9%, respectively.

Table A2: Effective sample size n_{eff} on the census healthcare data at selected budget levels $\rho = n_b/n$, averaged over 500 replications. Classical equals n_b by construction.

Method	$\rho = 0.05$	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.95$
`OptAPI` (Ours)	**32,270**	**62,230**	**140,200**	**225,700**	**310,300**
Confident Active Inference	30,010	58,090	129,000	208,900	308,700
Active Inference	30,360	58,500	129,200	209,100	308,900
PPI++	22,380	43,890	102,900	187,300	305,900
PPI	24,300	47,000	108,700	193,500	306,900
Classical	15,860	31,720	79,300	158,600	301,400

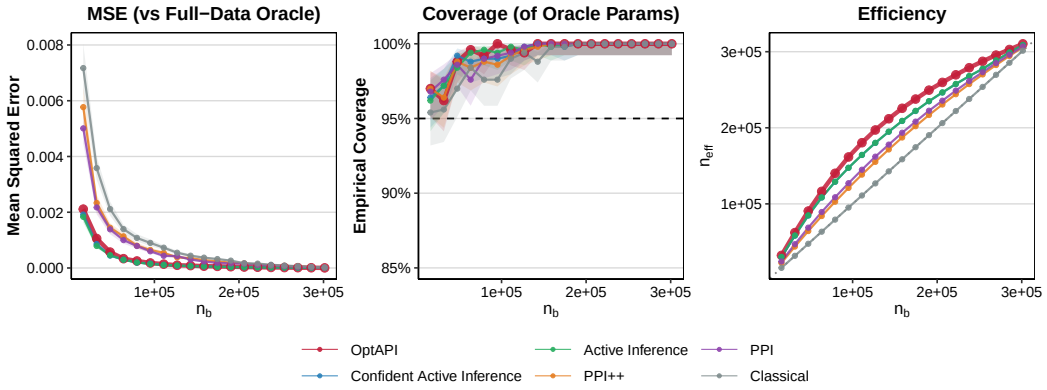


Figure A3: Real-world evaluation on the census healthcare dataset (logistic regression of a health-insurance indicator on income). `OptAPI` attains the largest effective sample size at every budget while maintaining valid coverage.

The two A-optimal methods attain slightly lower MSE than `OptAPI`. At $n_b/n = 50\%$, for example, the MSE is 6.48×10^{-5} for `OptAPI`, compared with 5.50×10^{-5} for Confident Active Inference and 5.72×10^{-5} for Active Inference. The difference is not significant at 5% but is significant against Confident Active Inference at 25% and 50%. This is the expected trade-off between the two criteria (Appendix E): MSE is a trace functional, which A-optimal designs minimize directly, whereas `OptAPI` minimizes the determinant. With $p = 2$ and no scale disparity between coordinates, targeting the trace costs the A-optimal designs relatively little volume, so they retain an edge on MSE while `OptAPI` retains the edge on joint confidence regions.

All methods attain coverage at or above the nominal level, ranging from 0.954 to 0.970 at $n_b/n = 5\%$ and reaching 1.000 for $n_b/n \geq 60\%$.[1]

G.5 Ablation Study: Mixing Parameter τ

The parameter $\tau \in [0, 1]$ controls the threshold for uniform mixing in our sampling policy, stabilizing the D-optimal active sampling probabilities. We conducted an ablation study utilizing the linear regression setup (locking the initial tuning parameter at $\lambda = 0.5$) to assess the framework's sensitivity to this choice.

Figure A4 illustrates the performance across $\tau \in \{0, 0.25, 0.5, 0.75, 1.0\}$. As expected, smaller values of τ (closer to pure active sampling) allow the policy to aggressively target high-leverage points, minimizing MSE and maximizing effective sample size. Conversely, setting $\tau = 1$ forces the design into a purely uniform sampling regime, drastically reducing efficiency and converging toward the Classical baseline. Intermediate values (e.g., $\tau = 0.25, 0.5$) offer a balanced trade-off, retaining the vast majority of the active efficiency gains while ensuring theoretical safeguards against pathological proxy models.

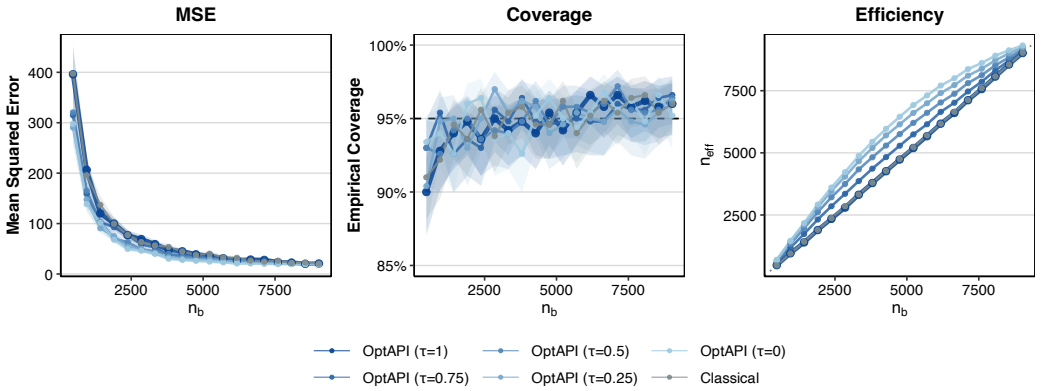


Figure A4: Ablation study on the uniform mixing parameter τ. Lower values encourage aggressive active sampling, maximizing efficiency, while higher values revert to safer, uniform sampling.

G.6 Ablation Study: Initial Proxy Weight λ

During the first phase of our algorithm, a pilot estimate of the optimal sampling distribution π must be constructed. Because the true optimal $\lambda^\star$ relies on the yet-to-be-sampled active data, we must initialize the procedure with a prior λ_{init}. We evaluate the sensitivity of our active sampling policy to this initialization by varying $\lambda_{\text{init}} \in \{0, 0.25, 0.5, 0.75, 1.0\}$.

[1] The over-coverage at large budgets is an artifact of the Full-Data Oracle rather than a defect of the procedure. Because $\theta^\star$ is the MLE of the same finite population being subsampled and the variance estimator carries no finite-population correction, $\widehat{\theta}$ approaches $\theta^\star$ as most of the pool is labeled, and the super-population intervals become conservative.

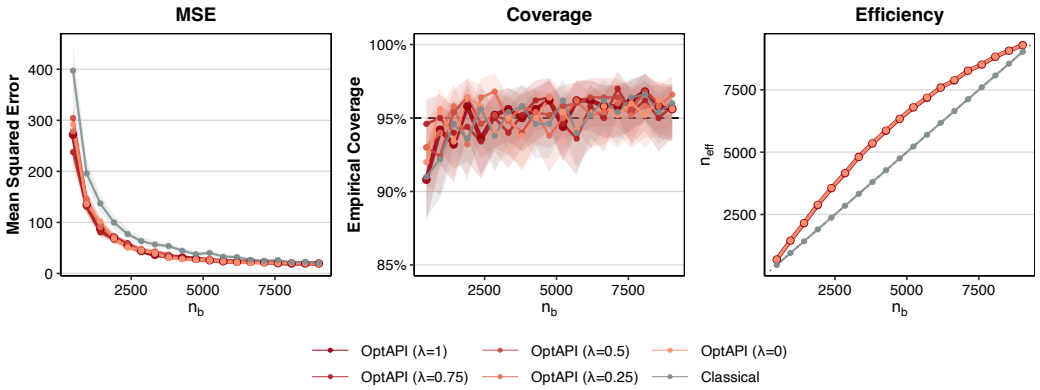


Figure A5: Ablation study on the initial proxy tuning parameter λ_{init}. The OptAPI framework remains highly efficient and strictly valid across all initialization strategies.

As shown in Figure A5, the framework is highly robust to λ initialization. Across all initializations, empirical coverage remains safely calibrated at the nominal 95% level. While initializing λ closer to the true empirical optimum yields the fastest convergence and highest initial efficiency, the dynamic power-tuning in the final phase successfully corrects for poor initializations, ensuring that the final estimator heavily outperforms uniform baselines regardless of the starting proxy weight.

G.7 Ablation Study: Pilot Sample Size n_l

`OptAPI` calibrates its design on the pilot set $\mathcal{D}_L$, which raises the question of how large n_l must be for the fitted surrogate to match a design built from the true nuisance. To isolate this effect, we hold the unlabeled pool and the label budget fixed and vary only the pilot size.

Setup. We use the heteroskedastic linear regression model of Section 4, with the same covariates, coefficients $\theta^\star = (1.0, 2.0, -1.5, 5.0)^\top$, heteroskedastic noise, and biased proxy. We fix the unlabeled pool at $n = 9{,}500$ and the budget at $n_b = 4{,}750$ (50% of the pool), and vary the pilot ratio $n_l/n_b \in \{0.02, 0.05, 0.10, 0.20, 0.45, 0.50\}$, i.e., $n_l \in \{95, 238, 475, 950, 2{,}138, 2{,}375\}$. Fresh data are drawn at each pilot ratio, so results are not paired across ratios. All other settings, including the surrogate construction, follow Appendix G.4, with 500 Monte Carlo replications per ratio. In addition to the six standard methods, we include an **Oracle** arm that runs `OptAPI` with the nuisance $v(x)$ fitted on all $n_l + 9{,}500$ points using the true labels, while keeping the pilot-based $\widetilde{\theta}_0$. `OptAPI` and the Oracle therefore differ only in the quality of the nuisance estimate, and the Oracle represents the $n_l \to \infty$ limit of the pilot-based surrogate.

Results. Figure A6 and Table A3 show that `OptAPI` reaches 99.3%–99.9% of the Oracle's n_{eff} at every pilot size, including $n_l = 95$, a pilot of only 2% of the budget and about 1% of the pool. The efficiency of `OptAPI` is essentially flat in n_l (6,303–6,385), corresponding to a gain of 32.7%–34.4% over the nominal budget. By contrast, every baseline sits at or below the nominal budget. The A-optimal methods again fall into the scale trap and are less efficient than Classical, and `OptAPI` improves on Confident Active Inference by about 38–41% and on PPI++ by about 34–35%. The confidence-ellipsoid volume $\mathcal{V}$ tells the same story: approximately 0.185 for both `OptAPI` and the Oracle, compared with 0.214 for Classical and 0.218 for Confident Active Inference.

Because the noise is heavy-tailed, MSE is noisy at 500 replications, and we interpret it qualitatively. The MSE of `OptAPI` ranges from 25.6 to 32.8 and that of the Oracle from 26.7 to 29.7, with overlapping confidence intervals at every ratio; both are roughly 25–30% below Classical (36.3–44.0). Confident Active Inference attains MSE comparable to or slightly below `OptAPI` at several ratios despite its much lower n_{eff}, reflecting once more the trace-versus-determinant distinction. Empirical coverage of the joint 95% ellipsoid lies between 0.934 and 0.966 for every method at every ratio. In summary, these results show that a small pilot suffices to calibrate the D-optimal design.

Table A3: Effective sample size n_{eff} in the pilot-size ablation with $n = 9{,}500$ and $n_b = 4{,}750$ fixed, averaged over 500 replications. The column `OptAPI`/Oracle reports the ratio of their n_{eff}.

n_l/n_b	n_l	**`OptAPI` (Ours)**	Oracle	`OptAPI`/Oracle	Confident Active Inference	Active Inference	PPI++	PPI	Classical
0.02	95	**6,303**	6,346	99.3%	4,470	3,967	4,667	4,138	4,750
0.05	238	**6,353**	6,364	99.8%	4,553	4,006	4,699	4,148	4,750
0.10	475	**6,334**	6,344	99.9%	4,544	4,000	4,702	4,160	4,750
0.20	950	**6,385**	6,399	99.8%	4,615	4,066	4,723	4,162	4,750
0.45	2,138	**6,380**	6,386	99.9%	4,599	4,073	4,759	4,151	4,750
0.50	2,375	**6,357**	6,369	99.8%	4,550	4,032	4,750	4,159	4,750

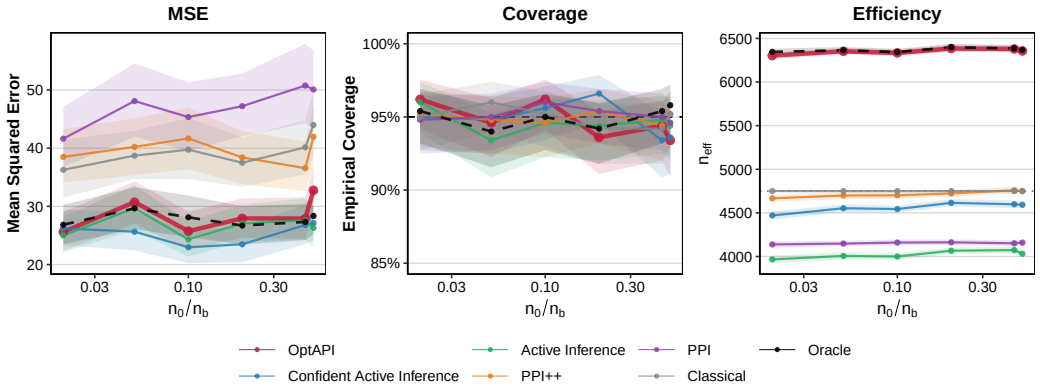


Figure A6: Ablation study on the pilot sample size n_l, with the pool $n = 9{,}500$ and budget $n_b = 4{,}750$ held fixed (horizontal axis n_l/n_b on a log scale; the reference line in the n_{eff} panel marks the nominal budget n_b). `OptAPI` matches the Oracle design even with $n_l = 95$.