General Bayesian Policy Learning

Masahiro Kato[1]

[1]Mizuho-DL Financial Technology Co., Ltd., Tokyo, Japan

Abstract

This study proposes a General Bayes framework for policy learning. We consider decision problems in which a decision-maker chooses an action from a given set to maximize expected welfare. Typical examples include treatment choice and portfolio optimization. In such problems, the statistical target is a decision rule, and predicting each potential outcome is not necessarily of primary interest. We formulate this policy-learning problem through loss-based Bayesian updating. Our main technical device is a squared-loss surrogate for welfare maximization. We show that maximizing empirical welfare over a policy class is equivalent to minimizing a scaled squared error in the outcome difference plus a quadratic penalty controlled by a tuning parameter $\zeta > 0$. The resulting General Bayes posterior over decision rules admits two equivalent characterizations: a Gaussian pseudo-likelihood representation and a decision-theoretic loss-based characterization. As one implementation, we introduce GBPLNet, a neural network implementation with a tanh-squashed output. Finally, we establish PAC-Bayes-type guarantees for the surrogate risk and the corresponding penalized welfare.

1 INTRODUCTION

Policy learning aims to train a policy function $\delta(x)$ that maps context features $x \in \mathcal{X}$ to an action a in a constrained action set, so as to maximize an expected outcome or, equivalently, to minimize an expected task loss. When the action set is discrete, many objectives can be written as a cost-sensitive 0-1 loss, where the loss penalizes choosing a suboptimal action with a weight given by an outcome gap. This perspective suggests using a learning rule that targets a task loss directly rather than fitting a potentially misspecified generative model.

General Bayes is a framework for updating beliefs using a loss function rather than a likelihood function [Bissiri et al., 2016]. Given a prior distribution and an empirical loss, General Bayes yields a generalized posterior that is coherent from a decision-theoretic viewpoint. This framework is attractive when a probabilistic model is misspecified, inconvenient, or unnecessary for the downstream task. This study develops a General Bayes framework for policy learning. The generalized posterior provides uncertainty over the decision rule itself. It can be used to summarize uncertainty in policy welfare and to assess whether the preferred action is stable across posterior draws. The goal is not to obtain uniform welfare improvements over empirical welfare maximization, weighted classification, or plug-in regression, but to retain a welfare-based objective while providing posterior information about policies and welfare.

A practical obstacle is that welfare objectives are typically linear in the policy, so they do not directly correspond to a convenient likelihood. We address this by rewriting welfare maximization as a squared-loss minimization problem. In the binary action case, the key surrogate loss has the form $\left(\frac{1}{\sqrt{\zeta}}\left(y\left(1\right) - y\left(0\right)\right) - \sqrt{\zeta} f\left(x\right)\right)^2$, where $f\left(x\right) \in [-1, 1]$ encodes a policy and $\zeta > 0$ is a tuning parameter. This surrogate admits a Gaussian pseudo-likelihood interpretation, so it enables a computationally convenient General Bayes posterior and standard approximation methods.

General Bayes allows any loss, so one could update directly using the negative welfare contribution as a loss. This direct update yields objectives that are linear in the policy. It does not induce the quadratic regularization in Theorem 4.1, and it does not provide a Gaussian pseudo-likelihood that can be exploited by standard Bayesian computation. The squared-loss surrogate yields a regression-style objective with stable gradients, an explicit regularization path indexed by ζ, and a unified extension to K actions and missing outcome settings.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2792–2827.

For K actions, a baseline-gap formulation in terms of outcome differences is natural, but it introduces a dependence on the baseline action through the regularization term. We therefore develop a baseline-free symmetric surrogate that operates on the full feedback vector. In missing outcome settings, for example, in observational studies and logged bandit feedback, outcome differences are not observed. We show that inverse propensity weighting (IPW) and doubly robust (DR) pseudo-outcomes yield explicit empirical losses that can be used for General Bayes updating, and we provide population-level target characterizations.

Contributions. We list our contributions as follows:

- We propose a General Bayes framework for policy learning that updates a prior over decision rules.
- For binary actions, we show that empirical welfare maximization is equivalent to minimizing a scaled squared-loss surrogate, up to a quadratic regularization controlled by ζ (Section 4).
- We define the associated General Bayes posterior, give its Gaussian pseudo-likelihood and loss-based characterizations, and discuss the role of the temperature parameter η (Section 4).
- For K actions, we present a baseline-gap surrogate and a baseline-free symmetric full-vector surrogate, and we discuss baseline choice (Section 5).
- For missing outcomes, we define empirical losses based on IPW and DR pseudo-outcomes, and we provide population-level target results (Section 7).
- We give PAC-Bayes bounds for the surrogate risk and translate them into bounds for penalized welfare (Section 8).
- We describe GBPLNet, neural networks with tanh-squashed outputs, as one implementation example for bounded scores (Section 6 and Appendix G).

We review related work in Appendix A. We refer to our framework as *General Bayesian Policy Learning (GBPL)* and our proposed implementation with neural networks as *GBPLNet*. Our proposed method can be applied to various tasks, including treatment choice [Kitagawa and Tetenov, 2018, Athey and Wager, 2021, Zhou et al., 2023] and portfolio optimization [Tallman and West, 2024a,b, Kato et al., 2024, Kato, 2024]. Note that we use empirical welfare maximization in a general sense, not restricted to counterfactual risk minimization [Swaminathan and Joachims, 2015, Kitagawa and Tetenov, 2018].

2 SETUP

Assume that there are K actions, indexed by $1, 2, \ldots, K$. Let $X \in \mathcal{X} \subseteq \mathbb{R}^k$ be a regressor and $Y(a) \in \mathcal{Y} \subseteq \mathbb{R}$ be an outcome for each action $a \in \{1, \ldots, K\}$. Let $Z = (X, Y(1), \ldots, Y(K))$ be a random vector following a distribution P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}^K$. Given i.i.d. observations $\mathcal{D} = \{z_i\}_{i=1}^n$ with $z_i = (x_i, y_i(1), \ldots, y_i(K))$, a decision-maker trains a decision rule $\delta\colon \mathcal{X} \to \{1, \ldots, K\}$ to maximize the expected welfare

$$V(\delta) = \mathbb{E}\left[\sum_{a\in\{1,\ldots,K\}} \mathbb{1}[\delta(X) = a] Y(a)\right]. \quad (1)$$

The Bayes optimal rule satisfies

$$\delta^\star(x) \in \arg\max_{a\in\{1,\ldots,K\}} \mathbb{E}[Y(a) \mid X = x]$$

for almost every x.

In many applications, outcomes are not fully observed, for example, under the potential outcome framework in causal inference and in bandit feedback in online learning. The missing outcome setting is introduced in Section 7. Sections 4 to 5 focus on the full feedback setting to isolate the surrogate construction.

3 BACKGROUND: GENERALIZED BAYESIAN UPDATING

Let Θ be a parameter space and let $\theta \in \Theta$ index a model for a decision rule or a score function. Let Π be a prior distribution on Θ. Given a loss function $\ell(\theta; z)$ and data $\mathcal{D} = \{z_i\}_{i=1}^n$, generalized Bayesian updating defines a generalized posterior

$$\mathrm{d}\Pi_\eta(\theta \mid \mathcal{D}) \propto \mathrm{d}\Pi(\theta) \exp\left(-\eta \sum_{i=1}^n \ell(\theta; z_i)\right), \quad (2)$$

where $\eta > 0$ is a temperature parameter.

Decision-theoretic derivation. A key result of Bissiri et al. [2016] is that (2) is the optimizer of a variational problem. Let Q be any probability distribution on Θ such that Q is absolutely continuous with respect to Π. Define

$$\mathcal{J}(Q) = \eta \mathbb{E}_{\theta\sim Q}\left[\sum_{i=1}^n \ell(\theta; z_i)\right] + D_{\mathrm{KL}}(Q \| \Pi). \quad (3)$$

Proposition 3.1. *Assume that the normalizing constant $Z_\eta(\mathcal{D}) = \int \exp\left(-\eta \sum_{i=1}^n \ell(\theta; z_i)\right) \mathrm{d}\Pi(\theta)$ is finite. Then the unique minimizer of $\mathcal{J}(Q)$ over all such Q is $Q = \Pi_\eta(\cdot \mid \mathcal{D})$.*

The proof is provided in Appendix B.

Relation to ordinary Bayes. If the loss is chosen as the negative log-likelihood, $\ell(\theta; z) = -\log p(z \mid \theta)$, and $\eta = 1$, then (2) reduces to ordinary Bayes' rule. For other losses, (2) yields a coherent update even when no global likelihood model is posited [Bissiri et al., 2016].

4 GBPL WITH BINARY ACTIONS

For simplicity, we first consider the binary action case, $K = 2$. We reindex actions 1 and 2 as 1 and 0. We consider a possibly randomized policy $\delta\colon \mathcal{X} \to [0,1]$. The expected welfare can be written as

$$V(\delta) = \mathbb{E}\left[\delta(X) Y(1) + (1-\delta(X)) Y(0)\right].$$

4.1 EMPIRICAL WELFARE MAXIMIZATION

Let $\mathcal{H}_{\mathrm{pol}}$ be a class of measurable maps $\delta\colon \mathcal{X} \to [0,1]$. The empirical welfare is

$$\widehat{V}(\delta) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta(X_i) Y_i(1) + (1-\delta(X_i)) Y_i(0)\right). \quad (4)$$

A natural estimator is

$$\widehat{\delta} \in \arg\max_{\delta \in \mathcal{H}_{\mathrm{pol}}} \widehat{V}(\delta). \quad (5)$$

From a Bayesian perspective, (5) does not specify a likelihood, so it is unclear how to define a posterior distribution over δ without further modeling choices.

4.2 SQUARED-LOSS SURROGATE

Define the score function $f(x) = 2\delta(x) - 1$ and the induced class $\mathcal{F}_{\mathcal{H}_{\mathrm{pol}}} = \{f\colon f(x) = 2\delta(x) - 1, \delta \in \mathcal{H}_{\mathrm{pol}}\}$. Let $\zeta > 0$ be a constant. We consider the squared-loss surrogate

$$\widehat{f}(\zeta) \in \quad (6)$$
$$\arg\min_{f \in \mathcal{F}_{\mathcal{H}_{\mathrm{pol}}}} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\sqrt{\zeta}}(Y_i(1) - Y_i(0)) - \sqrt{\zeta} f(X_i)\right)^2.$$

To state the equivalence, define the penalized welfare maximizer

$$\widetilde{\delta}(\lambda) \in \arg\max_{\delta \in \mathcal{H}_{\mathrm{pol}}}\left\{\widehat{V}(\delta) - \lambda\frac{1}{n}\sum_{i=1}^{n}(2\delta(X_i) - 1)^2\right\}. \quad (7)$$

Theorem 4.1. *For any $\zeta > 0$, a score $\widehat{f}(\zeta)$ minimizes (6) if and only if there exists a maximizer $\widetilde{\delta}(\zeta/4)$ of (7) such that $\widehat{f}(\zeta) = 2\widetilde{\delta}(\zeta/4) - 1$.*

The proof is provided in Appendix C. As $\zeta \to 0$, the penalty in (7) vanishes. If the maximizer in (5) is unique, then $\widetilde{\delta}(\lambda) \to \widehat{\delta}$ as $\lambda \to 0$. Without uniqueness, any limit point of $\widetilde{\delta}(\lambda)$ belongs to the set of maximizers in (5). The penalty term in (7) is minimized at $\delta(x) = 1/2$, so for finite ζ the surrogate induces shrinkage toward randomized policies.

4.3 LOSS AND GENERALIZED POSTERIOR

Let f_θ be a parametric score model indexed by $\theta \in \Theta$. In the binary action case, we define the loss as

$$\ell(\theta; (x, y(1), y(0))) = \frac{1}{2}\left(\frac{1}{\sqrt{\zeta}}(y(1) - y(0)) - \sqrt{\zeta} f_\theta(x)\right)^2. \quad (8)$$

The generalized posterior is then

$$\mathrm{d}\Pi_\eta(\theta \mid \mathcal{D}) \propto \mathrm{d}\Pi(\theta) \exp\left(-\eta\sum_{i=1}^{n}\ell(\theta; z_i)\right), \quad (9)$$

where $z_i = (x_i, y_i(1), y_i(0))$.

4.4 GAUSSIAN PSEUDO-LIKELIHOOD REPRESENTATION

Let $u = y(1) - y(0)$ denote the outcome difference. The exponential weight in (9) can be rewritten as

$$\exp(-\eta\ell(\theta; (x, y(1), y(0)))) = \exp\left(-\frac{\eta}{2\zeta}(u - \zeta f_\theta(x))^2\right).$$

Therefore, the exponential weight in (9) is identical to that induced by the Gaussian pseudo-likelihood

$$U \mid X = x, \quad \theta \sim \mathcal{N}(\zeta f_\theta(x), \zeta/\eta). \quad (10)$$

Equation (10) is an algebraic representation of the General Bayes update and is not an assumption that the true conditional distribution of U given X is Gaussian.

4.5 GENERAL BAYES INTERPRETATION

The posterior (9) has two equivalent characterizations. First, under the Gaussian pseudo-likelihood in (10), it has the form of an ordinary Bayesian posterior. Second, without committing to any likelihood, it is the unique optimizer of the variational problem in Proposition 3.1, so it is a coherent loss-based belief update [Bissiri et al., 2016].

The parameters ζ and η play different roles. The parameter ζ changes the learning objective itself, because it determines the strength of the quadratic regularization in Theorem 4.1. The parameter η changes posterior concentration, and it can be viewed as a calibration parameter. The parameter η also affects point estimators derived from the generalized posterior, for example, the MAP, because it rescales the empirical loss relative to the prior through (3). Under the Gaussian pseudo-likelihood in (10), ζ sets the scale of the conditional mean and ζ/η is the conditional variance, so ζ and η jointly determine the pseudo-likelihood scale. Selecting η is therefore a calibration problem, and practical strategies include information matching procedures [Lyddon et al., 2019] and adaptive learning rate procedures [Grünwald, 2012].

5 GBPL WITH MULTIPLE ACTIONS

We now extend the squared-loss surrogate and the posterior construction to K actions. We present two surrogate constructions. The first is based on outcome gaps relative to a baseline action. The second is baseline-free and symmetric, and it uses the full feedback vector.

5.1 BASELINE-GAP SURROGATE

Fix a baseline action K. Define outcome gaps $\overline{U}(a) = Y(a) - Y(K)$ for $a \in \{1, \dots, K-1\}$. For a possibly randomized policy $\delta\colon \mathcal{X} \to \Delta_K$, where $\Delta_K = \left\{p \in \mathbb{R}^K : p_a \geq 0, \sum_{a=1}^K p_a = 1\right\}$, welfare can be written as

$$V(\delta) = \mathbb{E}\left[Y(K) + \sum_{a=1}^{K-1} \delta_a(X)\, \overline{U}(a)\right].$$

The baseline term $\mathbb{E}[Y(K)]$ does not depend on δ.

Let $\mathcal{H}_{\mathrm{pol}}^{(K)}$ be a class of policies $\delta\colon \mathcal{X} \to \Delta_K$. For $a \in \{1, \dots, K-1\}$, define the componentwise score transform $f_a(x) = 2\delta_a(x) - 1$. Let $\mathcal{F}_{\mathcal{H}_{\mathrm{pol}}^{(K)}}$ denote the induced class of score vectors $f(x) = (f_1(x), \dots, f_{K-1}(x))$. For $\zeta > 0$, define the empirical surrogate

$$\widehat{f}^{\mathrm{Gap}}(\zeta) \in \arg\min_{f \in \mathcal{F}_{\mathcal{H}_{\mathrm{pol}}^{(K)}}} \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{K-1} \left(\frac{1}{\sqrt{\zeta}}(Y_i(a) - Y_i(K)) - \sqrt{\zeta} f_a(X_i)\right)^2. \tag{11}$$

Define the empirical welfare $\widehat{V}(\delta) = \frac{1}{n}\sum_{i=1}^n \sum_{a=1}^K \delta_a(X_i)\, Y_i(a)$ and the penalized empirical welfare maximizer

$$\widetilde{\delta}^{\mathrm{Gap}}(\lambda) \in \arg\max_{\delta \in \mathcal{H}_{\mathrm{pol}}^{(K)}} \left\{\widehat{V}(\delta) - \lambda \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{K-1} (2\delta_a(X_i) - 1)^2\right\}. \tag{12}$$

Theorem 5.1. *For any $\zeta > 0$, a score $\widehat{f}^{\mathrm{Gap}}(\zeta)$ minimizes (11) if and only if there exists a maximizer $\widetilde{\delta}^{\mathrm{Gap}}(\zeta/4)$ of (12) such that $\widehat{f}_a^{\mathrm{Gap}}(\zeta) = 2\widetilde{\delta}_a^{\mathrm{Gap}}(\zeta/4) - 1$ for every $a \in \{1, \dots, K-1\}$.*

The proof is provided in Appendix D.1.

5.2 BASELINE-FREE SYMMETRIC FULL-VECTOR SURROGATE

The baseline-gap surrogate is convenient, but its regularization depends on the chosen baseline through the term $\sum_{a=1}^{K-1} (2\delta_a(X) - 1)^2$. To remove baseline dependence, we use a full-vector surrogate that treats all actions symmetrically.

Let $\mathcal{H}_{\mathrm{pol}}^{(K)}$ be a class of policies $\delta\colon \mathcal{X} \to \Delta_K$. For $\zeta > 0$, define the empirical full-vector surrogate

$$\widehat{\delta}^{\mathrm{Full}}(\zeta) \in \arg\min_{\delta \in \mathcal{H}_{\mathrm{pol}}^{(K)}} \left\{\frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{K} \left(\frac{1}{\sqrt{\zeta}} Y_i(a) - \sqrt{\zeta}\delta_a(X_i)\right)^2\right\}. \tag{13}$$

Define the penalized empirical welfare maximizer

$$\widetilde{\delta}^{\mathrm{Full}}(\lambda) \in \arg\max_{\delta \in \mathcal{H}_{\mathrm{pol}}^{(K)}} \left\{\widehat{V}(\delta) - \lambda \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{K} \delta_a(X_i)^2\right\}. \tag{14}$$

The penalty term in (14) is minimized at $\delta_a(x) = 1/K$ for all a, so for finite ζ the full-vector surrogate induces shrinkage toward uniform randomization.

Theorem 5.2. *For any $\zeta > 0$, a policy $\widehat{\delta}^{\mathrm{Full}}(\zeta)$ minimizes (13) if and only if it maximizes (14) with $\lambda = \zeta/2$.*

The proof is provided in Appendix D.2.

5.3 BASELINE DEPENDENCE AND SYMMETRIC ALTERNATIVES

Theorems 5.1 and 5.2 show that both surrogate constructions correspond to penalized welfare maximization, but with different regularizers. In particular, the baseline-gap penalty depends on which action is chosen as baseline, while the full-vector penalty is baseline-free and symmetric across actions.

If the interest is the unregularized welfare maximizer, one may take ζ small so that the regularization is negligible. In that regime, baseline choice does not materially affect the maximizer, but it can affect numerical stability, because it changes the scaling and the variance of estimated gaps in missing outcome settings. If one prefers an invariant formulation for finite ζ, the full-vector surrogate provides a direct baseline-free alternative.

Another symmetric alternative is to average baseline-gap surrogates over baselines. This produces a pairwise gap objective that depends on all differences $Y(a) - Y(b)$. We do not pursue this direction further, because it introduces $K(K-1)/2$ pairwise terms.

5.4 SHIFT INVARIANCE AND CENTERING

Welfare is invariant to adding an action-common shift that depends on X, because $\sum_{a=1}^K \delta_a(X) = 1$.

That is, for any measurable $c(X)$, we have $\mathbb{E}\left[\sum_{a=1}^{K} \delta_a(X)(Y(a)+c(X))\right] = V(\delta) + \mathbb{E}[c(X)]$. The argmax over δ is therefore unchanged. The full-vector surrogate inherits the same invariance, because replacing $Y_i(a)$ by $Y_i(a)+c_i$ for any c_i that does not depend on a changes (13) only through additive constants, using again that $\sum_{a=1}^{K} \delta_a(X_i) = 1$. This observation provides a simple baseline-free variance reduction idea. One can center outcomes by an action-common baseline without changing the minimizer.

In the full feedback setting, a symmetric choice is $c_i = -\frac{1}{K}\sum_{a=1}^{K} Y_i(a)$, which centers the outcomes across actions for each unit. In missing outcome settings, one may use a regression estimate $c(x)$ of an action-common component of the outcome and replace pseudo-outcomes by centered pseudo-outcomes, which can reduce the variance of IPW and DR constructions without changing the optimal policy.

5.5 GENERAL BAYES POSTERIOR

Let $\delta_\theta(x) \in \Delta_K$ be a parametric policy model indexed by $\theta \in \Theta$. For the full-vector surrogate, define the loss

$$\ell_K^{\mathrm{Full}}(\theta; (x, y(1), \ldots, y(K))) \tag{15}$$
$$= \frac{1}{2}\sum_{a=1}^{K}\left(\frac{1}{\sqrt{\zeta}} y(a) - \sqrt{\zeta}\delta_{\theta,a}(x)\right)^2.$$

The generalized posterior is

$$\mathrm{d}\Pi_\eta(\theta \mid \mathcal{D}) \propto \mathrm{d}\Pi(\theta)\exp\left(-\eta\sum_{i=1}^{n}\ell_K^{\mathrm{Full}}(\theta; z_i)\right).$$

As in the binary case, (15) corresponds to a Gaussian pseudo-likelihood with conditionally independent components, $Y(a) \mid X = x, \theta \sim \mathcal{N}(\zeta\delta_{\theta,a}(x), \zeta/\eta)$ for $a \in \{1, \ldots, K\}$.

6 GBPLNET: AN IMPLEMENTATION EXAMPLE OF NEURAL NETWORKS

The generalized posterior is typically intractable for flexible models. Standard approximation approaches include MAP, Gaussian approximations, and stochastic gradient Langevin dynamics [Welling and Teh, 2011]. As one implementation example, we parameterize the bounded score $f_\theta(x) \in [-1, 1]$ by a neural network with a tanh-squashed output, $f_w(x) = \tanh(g_w(x))$. We refer to this implementation as the GBPLNet.

Given a fitted score, we form a policy by $\delta_w(x) = (f_w(x)+1)/2$. For evaluation in the full feedback setting, it is also convenient to use the deterministic decision rule $a_w(x) = \mathbb{1}[f_w(x) \geq 0]$. This deterministic rule selects the more likely action under the randomized policy $\delta_w(x)$. For K actions, a deterministic decision rule can be formed by $a_\theta(x) \in \arg\max_{a\in\{1,\ldots,K\}} \delta_{\theta,a}(x)$. MAP computation corresponds to minimizing the negative log generalized posterior, which is the empirical surrogate loss plus the negative log prior. Gaussian approximations and SGLD provide complementary ways to represent posterior uncertainty. Neural networks and tanh are not essential to the framework, and any model class that produces bounded scores can be used. Appendix G provides details.

Implementation summary The GBPL update is determined by a choice of empirical loss, a prior, and tuning parameters.

- Inputs are data, a parametric score or policy model, a prior Π, and tuning parameters ζ and η.
- In the full feedback setting, the loss is (8) for binary actions and (15) for K actions.
- In missing outcome settings, the loss is constructed as $\widehat{L}_n^{\mathrm{IPW}}(\theta)$ or $\widehat{L}_n^{\mathrm{DR}}(\theta)$.
- Outputs are the generalized posterior in (2) and derived decision rules, including a point estimate such as a MAP or approximate samples such as from SGLD.

7 GBPL WITH MISSING OUTCOMES

We now consider a missing outcome setting where only the outcome under the chosen action is observed. We focus on a standard observational study setting with bandit feedback.

7.1 OBSERVATION MODEL AND ASSUMPTIONS

We observe i.i.d. data $\{(X_i, A_i, Y_i)\}_{i=1}^{n}$, where $A_i \in \{1, \ldots, K\}$ is the chosen action and $Y_i = Y_i(A_i)$ is the observed outcome. Let $e_a(x) = \mathbb{P}(A = a \mid X = x)$ be the propensity score. We impose standard conditions [Rosenbaum and Rubin, 1983].

Assumption 7.1. ***Unconfoundedness.*** *The potential outcomes are conditionally independent of the action given the context,* $(Y(1), \ldots, Y(K)) \perp A \mid X$.

Assumption 7.2. ***Overlap.*** *There exists* $\epsilon > 0$ *such that* $e_a(X) \geq \epsilon$ *almost surely for all* $a \in \{1, \ldots, K\}$.

Under these assumptions, the welfare of a policy $\delta\colon \mathcal{X} \to \Delta_K$ satisfies $V(\delta) = \mathbb{E}\left[\sum_{a=1}^{K}\delta_a(X)\,\mathbb{E}[Y(a) \mid X]\right]$.

7.2 IPW PSEUDO-OUTCOMES

If the propensity score $e_a(x)$ is known, or replaced by an estimator $\widehat{e}_a(x)$, IPW constructs pseudo-outcomes

[Horvitz and Thompson, 1952]

$$\widetilde{Y}_i^{\mathrm{IPW}}(a) = \frac{\mathbb{1}[A_i = a] Y_i}{\widehat{e}_a(X_i)}. \tag{16}$$

Using the full-vector surrogate, we define the IPW empirical loss for a parametric policy δ_θ as $\widehat{L}_n^{\mathrm{IPW}}(\theta) = \sum_{i=1}^n \frac{1}{2} \sum_{a=1}^K \left(\frac{1}{\sqrt{\zeta}} \widetilde{Y}_i^{\mathrm{IPW}}(a) - \sqrt{\zeta} \delta_{\theta,a}(X_i) \right)^2$. The corresponding generalized posterior is $\mathrm{d}\Pi_\eta^{\mathrm{IPW}}(\theta \mid \mathcal{D}) \propto \mathrm{d}\Pi(\theta) \exp\left(-\eta \widehat{L}_n^{\mathrm{IPW}}(\theta)\right)$.

Theorem 7.1. *Assume that $\widehat{e}_a(X_i) > 0$ for every observation i and action a. Then minimizing $\widehat{L}_n^{\mathrm{IPW}}(\theta)$ over θ is equivalent to maximizing the IPW welfare estimator with a quadratic penalty,*

$$\frac{1}{n} \sum_{i=1}^n \sum_{a=1}^K \delta_{\theta,a}(X_i) \widetilde{Y}_i^{\mathrm{IPW}}(a) - \frac{\zeta}{2} \frac{1}{n} \sum_{i=1}^n \sum_{a=1}^K \delta_{\theta,a}(X_i)^2,$$

up to additive constants that do not depend on θ.

The proof is provided in Appendix E.1.

7.3 DR PSEUDO-OUTCOMES

Let $\gamma_a(x) = \mathbb{E}[Y(a) \mid X = x]$. Let $\widehat{\gamma}_a(x)$ be an estimator of $\gamma_a(x)$. The DR pseudo-outcome is given as follows [Bang and Robins, 2005]:

$$\widetilde{Y}_i^{\mathrm{DR}}(a) = \widehat{\gamma}_a(X_i) + \frac{\mathbb{1}[A_i = a](Y_i - \widehat{\gamma}_a(X_i))}{\widehat{e}_a(X_i)}. \tag{17}$$

Using the full-vector surrogate, define the DR empirical loss $\widehat{L}_n^{\mathrm{DR}}(\theta) = \sum_{i=1}^n \frac{1}{2} \sum_{a=1}^K \left(\frac{1}{\sqrt{\zeta}} \widetilde{Y}_i^{\mathrm{DR}}(a) - \sqrt{\zeta} \delta_{\theta,a}(X_i) \right)^2$. The generalized posterior is $\mathrm{d}\Pi_\eta^{\mathrm{DR}}(\theta \mid \mathcal{D}) \propto \mathrm{d}\Pi(\theta) \exp\left(-\eta \widehat{L}_n^{\mathrm{DR}}(\theta)\right)$.

In practice, $\widehat{\gamma}_a$ and $\widehat{e}_a$ may be estimated from the same data used to fit δ_θ. To reduce bias from the empirical process, we usually assume Donsker conditions for the nuisance estimators, or apply sample splitting, also called cross-fitting [Klaassen, 1987, van der Vaart, 2002]. See Chernozhukov et al. [2018], Zhou et al. [2023], and Schuler and van der Laan [2024] for details.

7.4 BINARY ACTION CASE

When $K = 2$, the surrogate in Section 4 is expressed in terms of the outcome difference $U = Y(1) - Y(0)$. In the missing outcome setting, U is not observed, but it can be replaced by an IPW or DR pseudo-difference.

For IPW, define $\widetilde{U}_i^{\mathrm{IPW}} = \frac{\mathbb{1}[A_i=1]Y_i}{\widehat{e}_1(X_i)} - \frac{\mathbb{1}[A_i=0]Y_i}{\widehat{e}_0(X_i)}$. For DR, define $\widetilde{U}_i^{\mathrm{DR}} = \left(\widehat{\gamma}_1(X_i) - \widehat{\gamma}_0(X_i)\right) + \frac{\mathbb{1}[A_i=1](Y_i - \widehat{\gamma}_1(X_i))}{\widehat{e}_1(X_i)} - \frac{\mathbb{1}[A_i=0](Y_i - \widehat{\gamma}_0(X_i))}{\widehat{e}_0(X_i)}$. Let $\widetilde{U}_i$ denote a generic pseudo-difference, with $\widetilde{U}_i = \widetilde{U}_i^{\mathrm{IPW}}$ for IPW and $\widetilde{U}_i = \widetilde{U}_i^{\mathrm{DR}}$ for DR. Let $f_\theta(x)$ be a score model with $f_\theta(x) \in [-1, 1]$ and define the binary loss by replacing $y(1) - y(0)$ in (8) with $\widetilde{U}_i$. For example, the DR empirical loss is $\widehat{L}_{n,\mathrm{bin}}^{\mathrm{DR}}(\theta) = \sum_{i=1}^n \frac{1}{2} \left(\frac{1}{\sqrt{\zeta}} \widetilde{U}_i^{\mathrm{DR}} - \sqrt{\zeta} f_\theta(X_i) \right)^2$, and the corresponding generalized posterior is defined as in (9) with $\widehat{L}_{n,\mathrm{bin}}^{\mathrm{DR}}$. The objective identity in Theorem 4.1 continues to hold after replacing U_i by $\widetilde{U}_i$, see Appendix F.4.

7.5 POPULATION TARGETS

The following proposition summarizes the population interpretation of IPW and DR pseudo-outcomes. Its proof is provided in Appendix E.2.

Proposition 7.2. *Assume Assumptions 7.1 and 7.2. Fix an action a.*

- *If $\widehat{e}_a = e_a$ almost surely, then $\mathbb{E}\left[\widetilde{Y}^{\mathrm{IPW}}(a) \mid X\right] = \gamma_a(X)$.*
- *If either $\widehat{e}_a = e_a$ almost surely or $\widehat{\gamma}_a = \gamma_a$ almost surely, then $\mathbb{E}\left[\widetilde{Y}^{\mathrm{DR}}(a) \mid X\right] = \gamma_a(X)$.*

Consequently, in either case the population minimizer of the full-vector squared-loss surrogate based on $\widetilde{Y}(a)$ matches the full feedback target, in the sense that the pointwise minimizer is the Euclidean projection of $(\gamma_1(x), \ldots, \gamma_K(x))/\zeta$ onto the simplex Δ_K.

7.6 STABILIZATION UNDER WEAK OVERLAP

Small estimated propensities can produce large IPW and DR pseudo-outcomes. In the additional experiments, we clip the binary propensity estimate as $\widehat{e}_1^{\mathrm{clip}}(x) = \max(\epsilon_{\mathrm{clip}}, \min(1 - \epsilon_{\mathrm{clip}}, \widehat{e}_1(x)))$ and set $\widehat{e}_0^{\mathrm{clip}}(x) = 1 - \widehat{e}_1^{\mathrm{clip}}(x)$. The algebraic equivalences in Theorems 4.1 and 7.1 continue to hold after replacing the pseudo-outcomes by their clipped versions, because these equivalences follow from expansion of the squared loss. The population target in Proposition 7.2 need not be preserved exactly after clipping. Appendix L reports synthetic and UCI experiments with logged feedback using the IPW and DR losses.

8 THEORETICAL ANALYSIS

This section summarizes theoretical guarantees for the surrogate loss and their implications for welfare. Proofs are provided in the Appendix.

8.1 POPULATION SURROGATE MINIMIZER

Binary case. In the binary action case, define $U = Y(1) - Y(0)$. For a measurable

$f\colon \mathcal{X} \to [-1,1]$, define the population surrogate risk $R_\zeta(f) = \mathbb{E}\left[\frac{1}{2}\left(\frac{1}{\sqrt{\zeta}}U - \sqrt{\zeta}f(X)\right)^2\right]$. Let $m(x) = \mathbb{E}[U \mid X = x]$. Then the unconstrained minimizer satisfies $f^\star(x) = m(x)/\zeta$. Under the constraint $f(x) \in [-1,1]$, the minimizer is the clipped conditional mean $f^\star(x) = \max(-1, \min(1, m(x)/\zeta))$. A formal statement and proof are given in Appendix F.1.

$K \geq 3$ **case.** For the full-vector surrogate, define $\gamma(x) = (\gamma_1(x), \ldots, \gamma_K(x))$ where $\gamma_a(x) = \mathbb{E}[Y(a) \mid X = x]$. Consider the population risk associated with (15). Conditionally on $X = x$, minimizing the conditional risk over $\delta \in \Delta_K$ yields the Euclidean projection of $\gamma(x)/\zeta$ onto Δ_K. A formal statement and proof are given in Appendix F.1.

8.2 PAC-BAYES BOUND UNDER A MOMENT CONDITION

We state a PAC-Bayes bound for the surrogate loss under a sub-exponential moment condition, following unbounded-loss PAC-Bayes analyses [Catoni, 2007, Haddouche et al., 2021, Alquier, 2024]. Let $\widehat{R}_\zeta(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; z_i)$ and $R_\zeta(\theta) = \mathbb{E}[\ell(\theta; Z)]$.

Assumption 8.1. *There exist constants $v > 0$ and $b > 0$ such that, for every $\theta \in \Theta$, the centered loss $\ell(\theta; Z) - R_\zeta(\theta)$ satisfies $\log \mathbb{E}[\exp(t(\ell(\theta; Z) - R_\zeta(\theta)))] \leq t^2 v/2$ for every t such that $|t| < 1/b$.*

Assumption 8.1 is a moment condition on the loss. In the full feedback binary setting with bounded scores, it holds under standard tail conditions on the outcome difference $U = Y(1) - Y(0)$. In missing outcome settings, the IPW and DR pseudo-outcomes in (16) and (17) involve inverse propensity weights, so the constants v and b can deteriorate as the overlap constant ϵ in Assumption 7.2 decreases. When nuisance estimators are learned from the same sample, the empirical losses are no longer i.i.d., and PAC-Bayes analyses typically rely on sample splitting or cross-fitting to control this dependence. In heavy-tailed regimes, one can replace the squared surrogate by a robust loss, for example, Catoni type or Huber type losses, and apply generalized Bayesian updating with the modified loss. Also see Catoni [2007].

Theorem 8.1. *Assume Assumption 8.1. Fix $\rho \in (0,1)$ and $t \in (0, 1/b)$. With probability at least $1 - \rho$ over $\mathcal{D}$, every probability distribution Q on Θ with $D_{\mathrm{KL}}(Q\|\Pi) < \infty$ satisfies $\mathbb{E}_{\theta\sim Q}[R_\zeta(\theta)] \leq \mathbb{E}_{\theta\sim Q}\left[\widehat{R}_\zeta(\theta)\right] + \frac{D_{\mathrm{KL}}(Q\|\Pi) + \log(1/\rho)}{tn} + \frac{tv}{2}$.*

A proof is provided in Appendix F.2.

Corollary 8.2. *Assume Assumption 8.1. Fix $\rho \in (0,1)$ and $t \in (0, 1/b)$. With probability at least $1 - \rho$ over $\mathcal{D}$, every probability distribution Q on Θ with $D_{\mathrm{KL}}(Q\|\Pi) < \infty$ satisfies $\left|\mathbb{E}_{\theta\sim Q}[R_\zeta(\theta)] - \mathbb{E}_{\theta\sim Q}\left[\widehat{R}_\zeta(\theta)\right]\right| \leq \frac{D_{\mathrm{KL}}(Q\|\Pi) + \log(2/\rho)}{tn} + \frac{tv}{2}$.*

A proof is provided in Appendix F.3.

8.3 WELFARE COROLLARIES

The squared-loss surrogate and penalized welfare are linked by an identity. The proofs in this subsection are provided in Appendices F.5–F.8.

Binary case. Define the population penalized welfare $W_\lambda(\delta) = V(\delta) - \lambda\mathbb{E}\left[(2\delta(X) - 1)^2\right]$. Let $\lambda = \zeta/4$ and let $f = 2\delta - 1$. Then $W_{\zeta/4}(\delta)$ differs from $R_\zeta(f)$ only through an additive constant that does not depend on δ. The next corollary makes the link explicit.

Corollary 8.3. *Let $\lambda = \zeta/4$. For any two policies δ_1, δ_2 in the binary action setting, let $f_j = 2\delta_j - 1$. Then, we have $W_\lambda(\delta_1) - W_\lambda(\delta_2) = \frac{1}{2}(R_\zeta(f_2) - R_\zeta(f_1))$. In particular, if $\delta_\lambda^\star$ maximizes W_λ over a class and $f_\zeta^\star$ minimizes R_ζ over the induced class, then we have $W_\lambda(\delta_\lambda^\star) - W_\lambda(\delta) = \frac{1}{2}\left(R_\zeta(f) - R_\zeta\left(f_\zeta^\star\right)\right)$.*

Corollary 8.4. *Let $\lambda = \zeta/4$. For any policy δ in the binary action setting, we have $W_\lambda(\delta) \leq V(\delta) \leq W_\lambda(\delta) + \lambda$. Consequently, if $\delta^\star$ maximizes V over a class and $\delta_\lambda^\star$ maximizes W_λ over the same class, then we have $V(\delta^\star) - V(\delta) \leq W_\lambda(\delta_\lambda^\star) - W_\lambda(\delta) + \lambda$.*

$K \geq 3$ **case.** For the full-vector surrogate, define the penalized welfare $W_\lambda^{\mathrm{Full}}(\delta) = V(\delta) - \lambda\mathbb{E}\left[\sum_{a=1}^K \delta_a(X)^2\right]$. The next corollary parallels Corollary 8.3.

Corollary 8.5. *Let $\lambda = \zeta/2$ and consider the full-vector surrogate risk associated with (15). For any two policies δ_1, δ_2, we have $W_\lambda^{\mathrm{Full}}(\delta_1) - W_\lambda^{\mathrm{Full}}(\delta_2) = \left(R_\zeta^{\mathrm{Full}}(\delta_2) - R_\zeta^{\mathrm{Full}}(\delta_1)\right)$, where $R_\zeta^{\mathrm{Full}}(\delta)$ is the population risk induced by (15) after substituting δ for δ_θ.*

Corollary 8.6. *Assume Assumptions 7.1 and 7.2. Consider the full-vector surrogate loss $\widehat{L}_n^{\mathrm{IPW}}(\theta)$ or $\widehat{L}_n^{\mathrm{DR}}$, and assume the conditional mean correctness property in Proposition 7.2, so that $\mathbb{E}\left[\widetilde{Y}(a) \mid X\right] = \gamma_a(X)$ for all a. Then the welfare identity in Corollary 8.5 continues to hold after replacing R_ζ^{Full} by the population risk defined using $\widetilde{Y}(a)$. In particular, for $\lambda = \zeta/2$ and any two policies δ_1, δ_2, we have $W_\lambda^{\mathrm{Full}}(\delta_1) - W_\lambda^{\mathrm{Full}}(\delta_2) = \left(\widetilde{R}_\zeta^{\mathrm{Full}}(\delta_2) - \widetilde{R}_\zeta^{\mathrm{Full}}(\delta_1)\right)$, where $\widetilde{R}_\zeta^{\mathrm{Full}}$ denotes the surrogate risk computed with $\widetilde{Y}(a)$.*

8.4 PAC-BAYES TO WELFARE

Combining Theorem 8.1 with Corollaries 8.3 and 8.5 yields welfare bounds. For example, in the binary case, we have $W_\lambda(\delta^\star_\lambda) - W_\lambda(\delta) = \frac{1}{2}\left(R_\zeta(f) - R_\zeta\left(f^\star_\zeta\right)\right)$. Thus, any upper bound on $R_\zeta(f)$ translates into a lower bound on penalized welfare, relative to the best-in-class penalized welfare. The same translation applies for missing outcome settings after replacing $Y(a)$ by IPW or DR pseudo-outcomes, because the optimization identities in Theorem 7.1 and Proposition 7.2 determine the population targets. These results concern penalized welfare. They do not characterize conditions under which GBPL uniformly outperforms empirical welfare maximization or plug-in methods in unpenalized welfare.

Table 1: Synthetic binary results, 100 trials. Columns report welfare mean, welfare variance, welfare standard error, regret mean, and regret standard error across trials.

method	welfare_mean	welfare_var	welfare_se	regret_mean	regret_se
DGP1					
DiffReg	1.0272	0.0081	0.0090	0.2291	0.0033
GBPLNet ($\zeta = 0.01$)	1.0240	0.0084	0.0091	0.2323	0.0034
GBPLNet ($\zeta = 0.1$)	1.0253	0.0081	0.0090	0.2310	0.0033
GBPLNet ($\zeta = 1.0$)	1.0298	0.0083	0.0091	0.2265	0.0034
GBPLNet (CV)	1.0295	0.0083	0.0091	0.2268	0.0033
PluginReg	1.0265	0.0085	0.0092	0.2298	0.0034
WeightedLogistic	1.0331	0.0083	0.0091	0.2232	0.0034
DGP2					
DiffReg	0.6810	0.0038	0.0061	0.3368	0.0029
GBPLNet ($\zeta = 0.01$)	0.7165	0.0040	0.0063	0.3013	0.0030
GBPLNet ($\zeta = 0.1$)	0.7192	0.0037	0.0061	0.2986	0.0030
GBPLNet ($\zeta = 1.0$)	0.7108	0.0038	0.0061	0.3070	0.0032
GBPLNet (CV)	0.7126	0.0037	0.0061	0.3052	0.0029
PluginReg	0.6908	0.0039	0.0063	0.3270	0.0032
WeightedLogistic	0.6746	0.0041	0.0064	0.3432	0.0034
DGP3					
DiffReg	0.4847	0.0046	0.0068	0.3019	0.0032
GBPLNet ($\zeta = 0.01$)	0.4847	0.0048	0.0070	0.3019	0.0031
GBPLNet ($\zeta = 0.1$)	0.4839	0.0049	0.0070	0.3026	0.0034
GBPLNet ($\zeta = 1.0$)	0.4875	0.0048	0.0069	0.2991	0.0032
GBPLNet (CV)	0.4878	0.0046	0.0068	0.2988	0.0032
PluginReg	0.4851	0.0047	0.0069	0.3015	0.0031
WeightedLogistic	0.4863	0.0045	0.0067	0.3003	0.0031

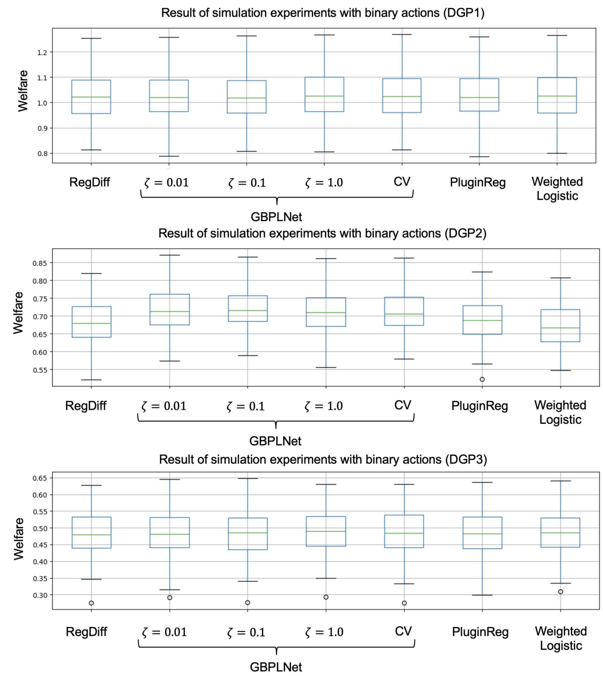

Figure 1: Synthetic binary welfare boxplots across 100 trials for DGP1, DGP2, and DGP3.

9 EXPERIMENTS

This section evaluates GBPLNet as an implementation example of the proposed General Bayes policy learning framework. Throughout, we focus on the full feedback setting, in which each observation contains the entire outcome vector $(Y(1), \ldots, Y(K))$. We report welfare and regret computed on held-out test data, and we aggregate results over 100 independent trials. In this section, we only show the results with binary actions using simulation datasets. The other detailed explanations and additional results are shown in Appendices J–K.

Evaluation protocol For each trial, we generate or load a dataset and split it into training, validation, and test sets. We train each method on the training set and select tuning parameters using the validation set. We then evaluate welfare on the test set using the empirical welfare formulas in (4) and their analogs for K actions. For binary actions, we report the realized test welfare of the deterministic policy induced by the learned score function f, $\widehat{\delta}(x) = \mathbb{1}[f(x) \geq 0]$. For K actions, we evaluate the welfare of the learned policy $\widehat{\delta}(x) \in \Delta_K$ by plugging $\widehat{\delta}(X_i)$ into (1) on the test set.

We define test regret relative to the oracle policy that maximizes the realized outcome for each test observation. In the binary case, the oracle action is $\delta^{\mathrm{orc}}(X_i) \in \arg\max_{a \in \{0,1\}} Y_i(a)$. We report the empirical regret on the test set, $\widehat{V}_{\text{test}}(\delta^{\mathrm{orc}}) - \widehat{V}_{\text{test}}\left(\widehat{\delta}\right)$. For each scenario and method, the tables report the mean, variance, and standard error of welfare across 100 trials, and the mean and standard error of regret across 100 trials.

We compare four values of the surrogate tuning parameter ζ for the proposed method. Three are fixed at $\zeta \in \{1, 0.1, 0.01\}$. One is selected by validation from the grid $\{1, 0.1, 0.01, 0.001\}$, and is reported as GBPLNet (CV).

For GBPLNet (CV), we select ζ by maximizing validation welfare rather than minimizing the validation surrogate loss. In the binary case, expanding the squared loss in (8) yields the term $\frac{1}{2\zeta}\frac{1}{n}\sum_{i=1}^{n} U_i^2$, where $U_i = Y_i(1) - Y_i(0)$, which does not depend on the fitted score, see Proposition G.1 in Appendix G. If the surrogate loss value itself is used for validation, this term can bias selection toward larger ζ values.

Synthetic Experiments We consider three synthetic data generating processes, denoted by DGP1, DGP2, and DGP3. Each trial uses $n = 5000$ observations and $d = 10$ features. Table 1 reports welfare and regret, and Figure 1 visualizes the distribution of welfare. We consider cases with $K = 2$. For the results with $K = 5$, see Appendix K.

The comparison methods are as follows. DiffReg fits a regression model for the outcome difference $Y(1) - Y(0)$ and chooses actions by its sign. PluginReg fits separate regression models for $Y(1)$ and $Y(0)$ and chooses the action with the larger predicted outcome. WeightedLogistic fits a weighted logistic classifier for the best action, with weights given by the observed outcome gap magnitude. GBPLNet trains the tanh transformed neural score model in Section 6 using the squared-loss surrogate.

Table 1 shows that relative performance is scenario dependent. In DGP1 and DGP3, the methods have similar welfare, while in DGP2 GBPLNet improves on DiffReg and PluginReg. WeightedLogistic attains the highest mean welfare in DGP1. The sensitivity to ζ is also scenario dependent.

10 CONCLUSION

This study proposes a General Bayes framework for policy learning (GBPL). The key technical device is a squared-loss surrogate that rewrites welfare maximization as a regression-style objective, with a Gaussian pseudo-likelihood interpretation. We demonstrate its implementation in both full-feedback and bandit-feedback settings. In our theoretical analysis, we give PAC-Bayes bounds for the surrogate risk and explicit corollaries for penalized welfare.

References

Pierre Alquier. User-friendly introduction to pac-bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, January 2024.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(236): 1–41, 2016.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37 (4):1591–1646, 2009.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2 edition, 1985.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Fischer Black and Robert Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43, 1992.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics lecture notes-monograph series. Institute of Mathematical Statistics, 2007.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, 2017.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2019. URL http://archive.ics.uci.edu/ml.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, 2011.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Peter Grünwald. The safe bayesian: learning the learning rate via the mixability gap. In *International Conference on Algorithmic Learning Theory (ALT)*, 2012.

Peter D. Grünwald and Nishant A. Mehta. Fast rates for general unbounded loss functions: From erm to generalized bayes. *Journal of Machine Learning Research*, 21 (56):1–80, 2020.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10): 1330, 2021.

Chris C. Holmes and Stephen G. Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Philippe Jorion. Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21 (3):279–292, 1986.

Masahiro Kato. General bayesian predictive synthesis, 2024. arXiv: 2406.09254.

Masahiro Kato, Kentaro Baba, Hibiki Kaibuchi, and Ryo Inokuchi. Bayesian portfolio optimization by predictive synthesis. In *IIAI International Congress on Advanced Applied Informatics (AAI)*, 2024.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Chris A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Annals of Statistics*, 15, 1987.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.

Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning (ICML)*, 2019.

Simon P. Lyddon, Chris C. Holmes, and Stephen G. Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

Charles F. Manski. Treatment choice under ambiguity induced by inferential problems. *Journal of Statistical Planning and Inference*, 105(1):67–82, 2002.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 1952.

Harry M Markowitz and G Peter Todd. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, volume 66. John Wiley & Sons, 2000.

David A. McAllester. Some PAC-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2 edition, 2007.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Alejandro Schuler and Mark van der Laan. Introduction to modern causal inference, 2024.

Matthias Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, March 2002. ISSN 1532-4435.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81, 2009.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(52):1731–1755, 2015.

Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.

Emily Tallman and Mike West. Bayesian predictive decision synthesis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):340–363, 04 2024a. ISSN 1369-7412.

Emily Tallman and Mike West. Predictive decision synthesis for portfolios: Betting on better models, 2024b. arXiv: 2405.01598.

Aad W. van der Vaart. Semiparametric statistics, 2002.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2): 49–60, 2013. doi: 10.1145/2641190.2641198. URL `https://doi.org/10.1145/2641190.2641198`.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2011.

Bryan Wilder, Eric Ewing, Bistra Dilkina, and Milind Tambe. End to end learning and optimization on graphs. In *Advances in Neural Information Processing Systems*, 2019.

Robert L. Winkler and Christopher B. Barry. A bayesian model for portfolio selection and revision. *The Journal of Finance*, 30(1):179–192, 1975.

Pei-Shien Wu and Ryan Martin. A Comparison of Learning Rate Selection Methods in Generalized Bayesian Inference. *Bayesian Analysis*, 18(1):105 – 132, 2023.

Baqun Zhang, Anastasios A. Tsiatis, Marie Davidian, Min Zhang, and Eric B. Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

A RELATED WORK

This appendix reviews related work and clarifies how the proposed GBPL framework is connected to existing lines of work. Our focus is on loss-based Bayesian updating and decision-theoretic learning rules for welfare objectives. Policy learning in causal inference and offline learning provides an important application area, but our framework is not tied to counterfactual semantics.

A.1 LOSS-BASED BAYES, GIBBS POSTERIORS, AND LEARNING RATE SELECTION

Generalized Bayesian updating, also called loss-based Bayes, replaces a likelihood by a task loss and yields a Gibbs-type posterior distribution [Bissiri et al., 2016]. This update is closely connected to the PAC-Bayes literature [McAllester, 1999, Seeger, 2002, Catoni, 2007] and to exponential weighting and aggregation methods for prediction under squared loss [Leung and Barron, 2006, Dalalyan and Tsybakov, 2008, Audibert, 2009]. From this viewpoint, the generalized posterior is an exponentially weighted distribution over decision rules, where the prior acts as a regularizer through a KL penalty [Bissiri et al., 2016].

A central practical issue is the choice of the learning rate, since the loss scale is not identified by probability axioms [Bissiri et al., 2016]. Several calibration principles have been developed, including the loss-likelihood bootstrap [Lyddon et al., 2019], generalized posterior calibration [Syring and Martin, 2019], and comparative studies of learning rate selection criteria in misspecified settings [Wu and Martin, 2023]. Safe Bayes is another data-driven approach that targets robustness to model misspecification [Grünwald, 2012].

When the loss is a negative log-likelihood, the generalized posterior becomes a tempered likelihood posterior, which connects General Bayes to power posteriors and fractional posteriors [Holmes and Walker, 2017, Bhattacharya et al., 2019]. For the squared-loss surrogate, the Gaussian pseudo-likelihood produces the same exponential weight as the General Bayes update, and Proposition 3.1 gives its loss-based variational characterization [Bissiri et al., 2016].

Finally, when exact sampling is infeasible, variational approximations provide tractable alternatives with theoretical support in the PAC-Bayes literature [Alquier et al., 2016]. Optimization-centric viewpoints also emphasize that generalized Bayesian procedures can be studied as regularized empirical risk minimization problems over distributions [Knoblauch et al., 2022]. For unbounded losses, including squared loss with unbounded outcomes, recent work develops fast-rate guarantees and learning-theoretic analyses for generalized Bayes under suitable moment conditions [Grünwald and Mehta, 2020].

A.2 BAYESIAN DECISION MAKING AND PREDICTIVE DECISION ANALYSIS

Bayesian inference is naturally connected to decision making through the Bayes act, which minimizes posterior expected loss, and through statistical decision theory more broadly [DeGroot, 1970, Berger, 1985, Robert, 2007]. This connection motivates Bayesian predictive decision rules in which decisions are evaluated under a predictive distribution and the loss encodes the target task. Recent work studies predictive decision analysis for welfare objectives in portfolio choice and other economic problems, including generalized Bayesian predictive decision rules [Tallman and West, 2024a,b, Kato, 2024].

A related line in machine learning and operations research is decision-focused learning, which trains predictive components to minimize a downstream decision objective. Representative examples include predict then optimize methods [Elmachtoub and Grigas, 2022, Donti et al., 2017, Wilder et al., 2019]. These works typically deliver point estimates of predictors or policies. Our focus is complementary, we construct a Gibbs posterior over decision rules via loss-based Bayesian updating.

A.3 PORTFOLIO CHOICE

Classical portfolio choice originates in mean-variance analysis [Markowitz, 1952, Markowitz and Todd, 2000]. Bayesian portfolio rules based on predictive distributions have also been studied, including early Bayesian treatments [Winkler and Barry, 1975] and influential empirical Bayes and shrinkage approaches [Jorion, 1986, Black and Litterman, 1992]. Recent work revisits portfolio choice from a predictive decision perspective and connects posterior predictive distributions to portfolio welfare, including generalized Bayesian formulations [Tallman and West, 2024b, Kato et al., 2024, Kato, 2024].

A.4 POLICY LEARNING WITH COUNTERFACTUALS AND OFFLINE FEEDBACK

In causal inference, treatment choice is a central application of welfare maximization [Manski, 2002, 2004, Stoye, 2009]. Policy learning methods train flexible decision rules using counterfactual identification strategies, including DR and IPW objectives [Dudík et al., 2011, Swaminathan and Joachims, 2015, Kitagawa and Tetenov, 2018, Athey and Wager, 2021, Zhou et al., 2023]. In offline learning from logged bandit feedback, counterfactual risk minimization is a canonical formulation, and Bayesian and PAC-Bayesian views have been developed [Swaminathan and Joachims, 2015, London and Sandler, 2019]. These literatures motivate practical empirical welfare objectives when outcomes are not fully observed. In contrast, our main development of the squared-loss surrogate and the generalized posterior does not rely on counterfactual interpretation, and it applies whenever a welfare objective can be written as an expectation under the data distribution.

A.5 CLASSIFICATION-BASED SURROGATES AND MULTI-ACTION EXTENSIONS

A large literature reduces welfare maximization to classification or regression by designing surrogate losses that target welfare directly. Outcome-weighted learning casts individualized treatment rules as weighted classification and motivates surrogate losses for tractable estimation [Zhao et al., 2012, Zhang et al., 2012]. Multi-action policy learning requires careful treatment of symmetry across actions, and baseline-gap formulations provide convenient reductions in several settings [Zhou et al., 2023]. Our baseline-free full-vector surrogate in Section 5 is motivated by the desire to retain symmetry across actions while keeping a squared-loss structure that supports Gaussian pseudo-likelihood computation.

A.6 PAC-BAYES BOUNDS AND GENERALIZATION GUARANTEES

PAC-Bayes bounds control the generalization error of randomized estimators and apply naturally to Gibbs posteriors [McAllester, 1999, Seeger, 2002, Catoni, 2007]. Recent work develops user-friendly PAC-Bayes bounds and extensions to unbounded losses under moment and tail conditions [Haddouche et al., 2021, Alquier, 2024, Grünwald and Mehta, 2020]. Our theoretical results in Section 8 follow this line and provide corollaries that translate surrogate risk bounds into welfare guarantees.

B PROOF OF PROPOSITION 3.1

Proof. Let $L(\theta) = \sum_{i=1}^{n} \ell(\theta; z_i)$ and let $Z_\eta(\mathcal{D}) = \int \exp(-\eta L(\theta))\, \mathrm{d}\Pi(\theta)$. Define $\Pi_\eta(\mathrm{d}\theta \mid \mathcal{D}) = \exp(-\eta L(\theta))\, Z_\eta(\mathcal{D})^{-1}\, \Pi(\mathrm{d}\theta)$. For any Q such that $Q \ll \Pi$,

$$D_{\mathrm{KL}}(Q \| \Pi_\eta(\cdot \mid \mathcal{D})) = \mathbb{E}_{\theta \sim Q}\left[\log\left(\frac{\mathrm{d}Q}{\mathrm{d}\Pi}\right)\right] + \eta \mathbb{E}_{\theta \sim Q}[L(\theta)] + \log Z_\eta(\mathcal{D}).$$

Rearranging gives

$$\mathcal{J}(Q) = D_{\mathrm{KL}}(Q \| \Pi_\eta(\cdot \mid \mathcal{D})) - \log Z_\eta(\mathcal{D}).$$

Since the last term does not depend on Q, $\mathcal{J}(Q)$ is minimized uniquely at $Q = \Pi_\eta(\cdot \mid \mathcal{D})$. □

C PROOF OF THEOREM 4.1

Proof. Write $U_i = Y_i(1) - Y_i(0)$. Expanding the objective in (6) gives

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\sqrt{\zeta}}U_i - \sqrt{\zeta} f(X_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\zeta}U_i^2 - 2U_i f(X_i) + \zeta f(X_i)^2\right).$$

The first term does not depend on f. Thus minimizing the surrogate is equivalent to maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\left(U_i f(X_i) - \frac{\zeta}{2} f(X_i)^2\right).$$

Substituting $f(x) = 2\delta(x) - 1$ and using (4) yields the penalized welfare criterion in (7) with $\lambda = \zeta/4$, up to constants that do not depend on δ.

Limit as $\zeta \to 0$. Let $\lambda_m \to 0$. For any $\delta \in \mathcal{H}_{\text{pol}}$,

$$\widehat{V}\left(\widetilde{\delta}\left(\lambda_m\right)\right) - \lambda_m \frac{1}{n}\sum_{i=1}^{n}\left(2\widetilde{\delta}\left(\lambda_m\right)\left(X_i\right) - 1\right)^2 \geq \widehat{V}\left(\delta\right) - \lambda_m \frac{1}{n}\sum_{i=1}^{n}\left(2\delta\left(X_i\right) - 1\right)^2.$$

Since $\left(2\delta\left(X_i\right) - 1\right)^2 \leq 1$, taking $m \to \infty$ yields

$$\liminf_{m\to\infty} \widehat{V}\left(\widetilde{\delta}\left(\lambda_m\right)\right) \geq \widehat{V}\left(\delta\right).$$

Therefore, any limit point of $\widetilde{\delta}\left(\lambda_m\right)$ is a maximizer of $\widehat{V}$. If the maximizer is unique, convergence follows. $\square$

D PROOFS IN SECTION 5 (POLICY LEARNING WITH MULTIPLE ACTIONS)

D.1 PROOF OF THEOREM 5.1

Proof. For each i and each $a \in \{1, \ldots, K-1\}$, let $U_i\left(a\right) = Y_i\left(a\right) - Y_i\left(K\right)$. Expanding the objective in (11) yields

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{K-1}\left(\frac{1}{\sqrt{\zeta}}U_i\left(a\right) - \sqrt{\zeta} f_a\left(X_i\right)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{K-1}\left(\frac{1}{\zeta}U_i\left(a\right)^2 - 2U_i\left(a\right) f_a\left(X_i\right) + \zeta f_a\left(X_i\right)^2\right).$$

Dropping constants and rescaling, minimizing is equivalent to maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{K-1}\left(U_i\left(a\right) f_a\left(X_i\right) - \frac{\zeta}{2} f_a\left(X_i\right)^2\right).$$

Substituting $f_a\left(x\right) = 2\delta_a\left(x\right) - 1$ yields the penalized welfare criterion in (12) with $\lambda = \zeta/4$, up to constants that do not depend on δ. $\square$

D.2 PROOF OF THEOREM 5.2

Proof. Expanding the objective in (13) yields

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{K}\left(\frac{1}{\sqrt{\zeta}}Y_i\left(a\right) - \sqrt{\zeta}\delta_a\left(X_i\right)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{K}\left(\frac{1}{\zeta}Y_i\left(a\right)^2 - 2Y_i\left(a\right)\delta_a\left(X_i\right) + \zeta\delta_a\left(X_i\right)^2\right).$$

Dropping the constant term and rescaling, minimizing is equivalent to maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{K}\left(Y_i\left(a\right)\delta_a\left(X_i\right) - \frac{\zeta}{2}\delta_a\left(X_i\right)^2\right),$$

which is (14) with $\lambda = \zeta/2$. $\square$

E PROOFS IN SECTION 7 (POLICY LEARNING WITH MISSING OUTCOMES)

E.1 PROOF OF THEOREM 7.1

Proof. Expanding $\widehat{L}_n^{\text{IPW}}\left(\theta\right)$ gives

$$\sum_{i=1}^{n}\frac{1}{2}\sum_{a=1}^{K}\left(\frac{1}{\sqrt{\zeta}}\widetilde{Y}_i^{\text{IPW}}\left(a\right) - \sqrt{\zeta}\delta_{\theta,a}\left(X_i\right)\right)^2 = \sum_{i=1}^{n}\frac{1}{2}\sum_{a=1}^{K}\left(\frac{1}{\zeta}\widetilde{Y}_i^{\text{IPW}}\left(a\right)^2 - 2\widetilde{Y}_i^{\text{IPW}}\left(a\right)\delta_{\theta,a}\left(X_i\right) + \zeta\delta_{\theta,a}\left(X_i\right)^2\right).$$

Dropping terms independent of θ yields the objective in Theorem 7.1. $\square$

E.2 PROOF OF PROPOSITION 7.2

Proof. Fix a. For IPW with $\widehat{e}_a = e_a$,

$$\mathbb{E}\left[\widetilde{Y}^{\text{IPW}}(a) \mid X\right] = \mathbb{E}\left[\frac{\mathbb{1}[A=a]\,Y(A)}{e_a(X)} \mid X\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}[A=a]\,Y(a)}{e_a(X)} \mid X, A\right] \mid X\right].$$

Under unconfoundedness, $\mathbb{E}[Y(a) \mid X, A] = \mathbb{E}[Y(a) \mid X] = \gamma_a(X)$. Thus the inner conditional expectation equals $\mathbb{1}[A=a]\,\gamma_a(X)/e_a(X)$ and taking expectation over $A \mid X$ yields $\gamma_a(X)$.

For DR, write

$$\widetilde{Y}^{\text{DR}}(a) = \widehat{\gamma}_a(X) + \frac{\mathbb{1}[A=a]\,(Y(a) - \widehat{\gamma}_a(X))}{\widehat{e}_a(X)}.$$

Taking conditional expectation given X and using unconfoundedness yields

$$\mathbb{E}\left[\widetilde{Y}^{\text{DR}}(a) \mid X\right] = \widehat{\gamma}_a(X) + \mathbb{E}\left[\frac{\mathbb{1}[A=a]}{\widehat{e}_a(X)}\,(\gamma_a(X) - \widehat{\gamma}_a(X)) \mid X\right].$$

If $\widehat{e}_a = e_a$, then $\mathbb{E}[\mathbb{1}[A=a]/\widehat{e}_a(X) \mid X] = 1$ and the right-hand side equals $\gamma_a(X)$. If $\widehat{\gamma}_a = \gamma_a$, then the second term is zero, and the right-hand side equals $\gamma_a(X)$. The final statement about the projection target follows from the conditional risk minimization argument in Appendix F.1. $\square$

F PROOFS IN SECTION 8 (THEORETICAL ANALYSIS)

F.1 POPULATION MINIMIZERS

Binary case.

Proof. For any fixed x, minimizing

$$\mathbb{E}\left[\frac{1}{2}\left(\frac{1}{\sqrt{\zeta}}U - \sqrt{\zeta}a\right)^2 \mid X = x\right] = \frac{1}{2\zeta}\mathbb{E}\left[(U - \zeta a)^2 \mid X = x\right]$$

over $a \in \mathbb{R}$ yields $a = m(x)/\zeta$ where $m(x) = \mathbb{E}[U \mid X = x]$. If a is constrained to $[-1, 1]$, the minimizer is the projection of $m(x)/\zeta$ onto $[-1, 1]$, which is $\max(-1, \min(1, m(x)/\zeta))$. $\square$

$K \geq 3$ case.

Proof. Fix x and consider the conditional risk for $\delta \in \Delta_K$,

$$\mathbb{E}\left[\frac{1}{2}\sum_{a=1}^{K}\left(\frac{1}{\sqrt{\zeta}}Y(a) - \sqrt{\zeta}\delta_a\right)^2 \mid X = x\right] = \frac{1}{2\zeta}\sum_{a=1}^{K}\mathbb{E}\left[(Y(a) - \zeta\delta_a)^2 \mid X = x\right].$$

Expanding the square yields a term that does not depend on δ and a quadratic term

$$\frac{1}{2\zeta}\sum_{a=1}^{K}(\gamma_a(x) - \zeta\delta_a)^2 = \frac{\zeta}{2}\sum_{a=1}^{K}(\delta_a - \gamma_a(x)/\zeta)^2.$$

Thus the minimizer is the Euclidean projection of $\gamma(x)/\zeta$ onto Δ_K. $\square$

F.2 PAC-BAYES BOUND

Proof. Fix $t \in (0, 1/b)$. For each $\theta \in \Theta$, Assumption 8.1 applied with $-t$ gives

$$\mathbb{E}_{\mathcal{D}}\left[\exp\left(tn\left(R_{\zeta}(\theta) - \widehat{R}_{\zeta}(\theta)\right)\right)\right] \leq \exp\left(\frac{nt^2 v}{2}\right).$$

Integrating this inequality with respect to Π and applying Fubini's theorem gives

$$\mathbb{E}_{\mathcal{D}}\left[\int \exp\left(tn\left(R_{\zeta}(\theta) - \widehat{R}_{\zeta}(\theta)\right)\right) \mathrm{d}\Pi(\theta)\right] \leq \exp\left(\frac{nt^2 v}{2}\right).$$

Markov's inequality implies that, with probability at least $1-\rho$,

$$\int \exp\left(tn\left(R_{\zeta}(\theta) - \widehat{R}_{\zeta}(\theta)\right)\right) \mathrm{d}\Pi(\theta) \leq \frac{1}{\rho}\exp\left(\frac{nt^2 v}{2}\right).$$

On this event, the variational inequality for KL divergence gives, for every Q with $D_{\mathrm{KL}}(Q\|\Pi) < \infty$,

$$tn\mathbb{E}_{\theta\sim Q}\left[R_{\zeta}(\theta) - \widehat{R}_{\zeta}(\theta)\right] \leq D_{\mathrm{KL}}(Q\|\Pi) + \log\int \exp\left(tn\left(R_{\zeta}(\theta) - \widehat{R}_{\zeta}(\theta)\right)\right) \mathrm{d}\Pi(\theta).$$

Substituting the preceding bound and dividing by tn yields the result. □

F.3 PROOF OF COROLLARY 8.2

Proof. Apply Theorem 8.1 to ℓ with failure probability $\rho/2$ and the fixed parameter t. This bounds $\mathbb{E}_{\theta\sim Q}[R_{\zeta}(\theta)] - \mathbb{E}_{\theta\sim Q}\left[\widehat{R}_{\zeta}(\theta)\right]$. Apply the same theorem to $-\ell$ with failure probability $\rho/2$. Assumption 8.1 also holds for $-\ell$ because it is stated for every t with $|t| < 1/b$. This yields an upper bound on $\mathbb{E}_{\theta\sim Q}\left[\widehat{R}_{\zeta}(\theta)\right] - \mathbb{E}_{\theta\sim Q}[R_{\zeta}(\theta)]$. A union bound completes the proof. □

F.4 PROOF FOR BINARY MISSING-OUTCOME OBJECTIVES

Proof. The proof of Theorem 4.1 uses only algebraic expansion of the squared loss and does not rely on U_i being the true outcome difference. Therefore the same expansion yields the same optimization identity after replacing U_i by any scalar pseudo-difference $\widetilde{U}_i$, including IPW and DR estimators. □

F.5 PROOF OF COROLLARY 8.3

Proof. Let $f = 2\delta - 1$. Expanding $R_{\zeta}(f)$ yields

$$R_{\zeta}(f) = \frac{1}{2\zeta}\mathbb{E}\left[U^2\right] - \mathbb{E}[Uf(X)] + \frac{\zeta}{2}\mathbb{E}\left[f(X)^2\right].$$

Since $Uf = 2U\delta - U$, we have

$$R_{\zeta}(f) = \frac{1}{2\zeta}\mathbb{E}\left[U^2\right] - 2\mathbb{E}[U\delta(X)] + \mathbb{E}[U] + \frac{\zeta}{2}\mathbb{E}\left[(2\delta(X)-1)^2\right].$$

Rearranging yields

$$\mathbb{E}[U\delta(X)] - \frac{\zeta}{4}\mathbb{E}\left[(2\delta(X)-1)^2\right] = C - \frac{1}{2}R_{\zeta}(f),$$

where C depends on $\mathbb{E}[U]$ and $\mathbb{E}\left[U^2\right]$ but not on δ. Adding $\mathbb{E}[Y(0)]$ on both sides gives $W_{\zeta/4}(\delta) = C' - R_{\zeta}(f)/2$ for a constant C' that does not depend on δ. Taking differences between δ_1 and δ_2 cancels the constant. □

F.6 PROOF OF COROLLARY 8.4

Proof. Since $0 \le (2\delta(X) - 1)^2 \le 1$, we have $0 \le \mathbb{E}\left[(2\delta(X) - 1)^2\right] \le 1$. Therefore,

$$W_\lambda(\delta) = V(\delta) - \lambda\mathbb{E}\left[(2\delta(X) - 1)^2\right] \le V(\delta) \le W_\lambda(\delta) + \lambda.$$

The final inequality follows by applying the first inequality to $\delta_\lambda^\star$ and the second inequality to $\delta^\star$ and taking differences. $\square$

F.7 PROOF OF COROLLARY 8.5

Proof. Expanding the full-vector population risk yields

$$R_\zeta^{\text{Full}}(\delta) = \frac{1}{2\zeta}\sum_{a=1}^{K}\mathbb{E}\left[Y(a)^2\right] - \sum_{a=1}^{K}\mathbb{E}\left[Y(a)\,\delta_a(X)\right] + \frac{\zeta}{2}\mathbb{E}\left[\sum_{a=1}^{K}\delta_a(X)^2\right].$$

The welfare is $V(\delta) = \sum_{a=1}^{K}\mathbb{E}\left[Y(a)\,\delta_a(X)\right]$. Thus,

$$W_{\zeta/2}^{\text{Full}}(\delta) = V(\delta) - \frac{\zeta}{2}\mathbb{E}\left[\sum_{a=1}^{K}\delta_a(X)^2\right] = C - R_\zeta^{\text{Full}}(\delta),$$

where $C = \frac{1}{2\zeta}\sum_{a=1}^{K}\mathbb{E}\left[Y(a)^2\right]$ is a constant that does not depend on δ. Taking differences yields the result. $\square$

F.8 PROOF OF COROLLARY 8.6

Proof. Let $\widetilde{R}_\zeta^{\text{Full}}(\delta)$ denote the population risk associated with $\widehat{L}_n^{\text{DR}}$ after substituting a deterministic policy δ for δ_θ and taking expectation. Expanding the square yields

$$\widetilde{R}_\zeta^{\text{Full}}(\delta) = C - \sum_{a=1}^{K}\mathbb{E}\left[\widetilde{Y}(a)\,\delta_a(X)\right] + \frac{\zeta}{2}\mathbb{E}\left[\sum_{a=1}^{K}\delta_a(X)^2\right],$$

where $C = \frac{1}{2\zeta}\sum_{a=1}^{K}\mathbb{E}\left[\widetilde{Y}(a)^2\right]$ does not depend on δ. Under the conditional mean correctness property in Proposition 7.2,

$$\mathbb{E}\left[\widetilde{Y}(a)\,\delta_a(X)\right] = \mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}(a) \mid X\right]\delta_a(X)\right] = \mathbb{E}\left[\gamma_a(X)\,\delta_a(X)\right] = \mathbb{E}\left[Y(a)\,\delta_a(X)\right].$$

Thus, up to a constant, $\widetilde{R}_\zeta^{\text{Full}}(\delta)$ differs from $W_{\zeta/2}^{\text{Full}}(\delta)$ only through a sign, and taking differences between δ_1 and δ_2 yields the stated identity. $\square$

G DETAILS OF THE GBPLNET

This appendix summarizes computational details for the GBPLNet, the neural network implementation example introduced in Section 6. We focus on the binary action setting with full feedback and comment on extensions.

G.1 BINARY ACTION SETTING AND THE GBPLNET SCORE MODEL

We work with i.i.d. data $\mathcal{D} = \{Z_i\}_{i=1}^n$ where $Z_i = (X_i, Y_i(1), Y_i(0))$. Let $U_i = Y_i(1) - Y_i(0)$. The GBPLNet models a bounded score function by $f_w(x) = \tanh(g_w(x))$, where $g_w(x)$ is an unconstrained neural network mapping $\mathcal{X}$ to $\mathbb{R}$ and $w \in \mathcal{W}$ denotes its weights. By construction, $f_w(x) \in (-1, 1)$ for all x.

The corresponding policy is $\delta_w(x) = (f_w(x) + 1)/2$. When a deterministic policy is required, we use the thresholding rule $a_w(x) = \mathbb{1}[f_w(x) \ge 0]$.

G.2 GENERAL BAYES POSTERIOR AND MAP OBJECTIVE

The GBPLNet substitutes f_w into the binary surrogate loss in (8),

$$\ell\left(w;\left(x,y\left(1\right),y\left(0\right)\right)\right)=\frac{1}{2}\left(\frac{1}{\sqrt{\zeta}}\left(y\left(1\right)-y\left(0\right)\right)-\sqrt{\zeta}f_w\left(x\right)\right)^2.$$

The generalized posterior is

$$\mathrm{d}\Pi_\eta\left(w\mid\mathcal{D}\right)\propto\mathrm{d}\Pi\left(w\right)\exp\left(-\eta\sum_{i=1}^{n}\ell\left(w;Z_i\right)\right).$$

In many implementations, Π is chosen as an isotropic Gaussian prior over weights, $\Pi=\mathcal{N}\left(0,\tau^2 I\right)$. Then $-\log\Pi\left(w\right)=\frac{1}{2\tau^2}\sum_{j=1}^{p}w_j^2+\mathrm{const}$, where p is the number of parameters. The MAP estimator satisfies

$$\widehat{w}_{\mathrm{MAP}}\in\arg\min_{w\in\mathcal{W}}\left\{\eta\sum_{i=1}^{n}\ell\left(w;Z_i\right)+\frac{1}{2\tau^2}\sum_{j=1}^{p}w_j^2\right\}.$$

This objective is standard neural network training under the squared surrogate loss with weight decay, up to the scaling by η.

G.3 OBJECTIVE DECOMPOSITION

The binary surrogate loss admits a simple decomposition, which is useful for implementation and for interpreting the effect of ζ.

Proposition G.1. *Let $U_i=Y_i\left(1\right)-Y_i\left(0\right)$. For any $\zeta>0$ and any score function f, we have*

$$\frac{1}{2}\left(\frac{1}{\sqrt{\zeta}}U_i-\sqrt{\zeta}f\left(X_i\right)\right)^2=\frac{1}{2\zeta}U_i^2-U_if\left(X_i\right)+\frac{\zeta}{2}f\left(X_i\right)^2.$$

Consequently,

$$\sum_{i=1}^{n}\frac{1}{2}\left(\frac{1}{\sqrt{\zeta}}U_i-\sqrt{\zeta}f\left(X_i\right)\right)^2=\frac{1}{2\zeta}\sum_{i=1}^{n}U_i^2-\sum_{i=1}^{n}U_if\left(X_i\right)+\frac{\zeta}{2}\sum_{i=1}^{n}f\left(X_i\right)^2.$$

Proposition G.1 implies that, for a fixed ζ, minimizing the surrogate empirical loss is equivalent to minimizing $-\sum_{i=1}^{n}U_if\left(X_i\right)+\frac{\zeta}{2}\sum_{i=1}^{n}f\left(X_i\right)^2$, up to an additive constant that does not depend on f. This is consistent with Theorem 4.1 after the reparameterization $f=2\delta-1$.

When tuning ζ by a validation criterion, Proposition G.1 clarifies that the full squared loss contains the term $\frac{1}{2\zeta}\sum_{i=1}^{n}U_i^2$ that does not depend on the fitted score. If the validation criterion is taken as the surrogate loss itself, this term can bias selection toward larger ζ values. A practical alternative is to tune ζ by validation welfare, or by the reduced objective that drops the constant term.

G.4 GAUSSIAN APPROXIMATION

A Gaussian approximation to the generalized posterior is obtained by a second order expansion around a point estimator, such as the MAP. Let $\widehat{w}_{\mathrm{MAP}}$ be a MAP estimate and define

$$U\left(w\right)=\eta\sum_{i=1}^{n}\ell\left(w;Z_i\right)-\log\Pi\left(w\right).$$

Let H be the Hessian of U at $\widehat{w}_{\mathrm{MAP}}$. A Gaussian approximation takes the form $w\mid\mathcal{D}\approx\mathcal{N}\left(\widehat{w}_{\mathrm{MAP}},H^{-1}\right)$. In large networks, H is high dimensional, so implementations typically use diagonal, block diagonal, or low rank approximations.

G.5 STOCHASTIC GRADIENT LANGEVIN DYNAMICS

SGLD provides an approximate sampling method for generalized posteriors. Let $\widehat{U}_t(w)$ be a minibatch approximation of $U(w)$ based on a batch $\mathcal{B}_t \subset \{1, \ldots, n\}$, scaled to match the full sample objective. Given a stepsize sequence $\{\epsilon_t\}$, SGLD iterates

$$w_{t+1} = w_t - \frac{\epsilon_t}{2} \nabla \widehat{U}_t(w_t) + \sqrt{\epsilon_t}\xi_t, \quad \xi_t \sim \mathcal{N}(0, I).$$

After a burn in period, draws are collected by thinning. In practice, gradient clipping is often helpful when the outcomes have heavy tails, because the squared loss can yield large gradients.

G.6 POSTERIOR PREDICTIVE WELFARE AND CREDIBLE INTERVALS

Let $\{w^{(s)}\}_{s=1}^{S}$ be draws from the generalized posterior, obtained for example by SGLD. For each draw, we can evaluate a policy on a held out sample $\{Z_i\}_{i=1}^{n}$.

If we use the deterministic decision rule $a_w(x) = \mathbb{1}[f_w(x) \geq 0]$, the welfare draw is

$$\widehat{V}^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \Big(a_{w^{(s)}}(X_i) Y_i(1) + (1 - a_{w^{(s)}}(X_i)) Y_i(0) \Big).$$

If we instead report welfare for the randomized policy $\delta_w(x) = (f_w(x) + 1)/2$, then

$$\widehat{V}^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \Big(\delta_{w^{(s)}}(X_i) Y_i(1) + (1 - \delta_{w^{(s)}}(X_i)) Y_i(0) \Big).$$

A 95% credible interval for welfare is given by the empirical quantiles of $\left\{\widehat{V}^{(s)}\right\}_{s=1}^{S}$, the 0.025 and 0.975 quantiles. These intervals summarize posterior uncertainty under the generalized posterior, and their frequentist coverage depends on calibration of the loss scale.

G.7 MULTI-ACTION EXTENSION

For K actions, the full-vector surrogate in (15) requires a policy model $\delta_\theta(x) \in \Delta_K$. A neural implementation replaces δ_θ by a network that outputs logits $h_w(x) \in \mathbb{R}^K$ and maps them to Δ_K by the softmax function. This multi-action implementation is symmetric across actions, and it avoids baseline dependence in the surrogate construction.

G.8 MINIMAL PYTORCH IMPLEMENTATION FOR THE GBPLNET

The following code illustrates MAP training for the binary surrogate with a tanh-squashed neural network score.

```
import torch
import torch.nn as nn

class TanhScoreNet(nn.Module):
    def __init__(self, in_dim, hidden=(128, 128)):
        super().__init__()
        layers = []
        d = in_dim
        for h in hidden:
            layers += [nn.Linear(d, h), nn.ReLU()]
            d = h
        layers += [nn.Linear(d, 1)]
        self.net = nn.Sequential(*layers)

    def forward(self, x):
        return torch.tanh(self.net(x)).view(-1)
```

```

def surrogate_loss(y1, y0, f, zeta):
    d = y1 - y0
    return 0.5 * (d / (zeta ** 0.5) - (zeta ** 0.5) * f) ** 2

def map_objective(model, xb, y1b, y0b, n_total, eta, tau, zeta):
    f = model(xb)
    loss_batch = surrogate_loss(y1b.view(-1), y0b.view(-1), f, zeta).sum()
    B = xb.size(0)
    data_term = eta * (n_total / B) * loss_batch
    prior_term = 0.0
    for p in model.parameters():
        prior_term = prior_term + (p ** 2).sum()
    prior_term = 0.5 * prior_term / (tau ** 2)
    return data_term + prior_term
```

H POSTERIOR VISUALIZATION EXAMPLE

This section illustrates how the generalized posterior in GBPLNet (Section 6) induces uncertainty bands for the score function and credible intervals for welfare, under a controlled full feedback binary setting.

H.1 DATA GENERATING PROCESS AND TRAINING SETUP

We generate one-dimensional covariates $X \in \mathbb{R}$ with $X \sim \text{Unif}\,(-2.5, 2.5)$. The baseline mean function is $\gamma_0\,(x) = 0.2x + 0.2\sin\,(1.5x)$ and the conditional outcome gap is $\tau\,(x) = 1.2\sin\,(x)$. We generate potential outcomes as $Y\,(0) = \gamma_0\,(X) + \epsilon_0$ and $Y\,(1) = \gamma_0\,(X) + \tau\,(X) + \epsilon_1$, with independent $\epsilon_0, \epsilon_1 \sim \mathcal{N}\left(0, 0.6^2\right)$. We set $\zeta = 1.0$ to avoid trivial saturation of the bounded score in most of the covariate range. The sample size is $n = 1500$, and we split the data into training, validation, and test sets with proportions 0.6, 0.2, and 0.2.

For the score model, we use the GBPLNet parameterization $f_w\,(x) = \tanh\,(g_w\,(x))$ with a two hidden layer MLP g_w of widths $(64, 64)$ and ReLU activation. We compute a MAP estimate under the squared-loss surrogate with a Gaussian prior $\Pi = \mathcal{N}\left(0, \tau^2 I\right)$, using $\tau = 1.0$ and $\eta = 1.0$. We train with minibatches of size 128, learning rate 10^{-3}, and weight decay 10^{-4}, and we apply early stopping based on the validation surrogate loss. In this run, the best validation surrogate loss is 0.3482.

H.2 GENERALIZED POSTERIOR SAMPLING BY SGLD

To visualize the generalized posterior, we approximate draws from $\Pi_\eta\,(\cdot \mid \mathcal{D})$ by SGLD starting from the MAP solution. We use burn in 1200 iterations, collect 300 draws, thin by a factor 8, and use step size 2×10^{-5} with batch size 128. These settings are chosen for a stable qualitative illustration rather than for asymptotic accuracy.

H.3 POSTERIOR DRAWS FOR THE SCORE FUNCTION

Figure 2 plots posterior draws of $f_w\,(x)$ on a grid, together with the posterior mean and a pointwise 95% credible band. The figure also overlays the population target $\max\,(-1, \min\,(1, \tau\,(x)\,/\zeta))$ from Section 8, which equals $\max\,(-1, \min\,(1, 1.2\sin\,(x)))$ for $\zeta = 1.0$. The posterior mean tracks the target in regions where the target lies in $(-1, 1)$ and saturates near the boundaries when the target exceeds the action range. The credible band widens in transition regions where the decision boundary $f_w\,(x) = 0$ is sensitive to the fit, which is reflected in posterior variability in downstream welfare.

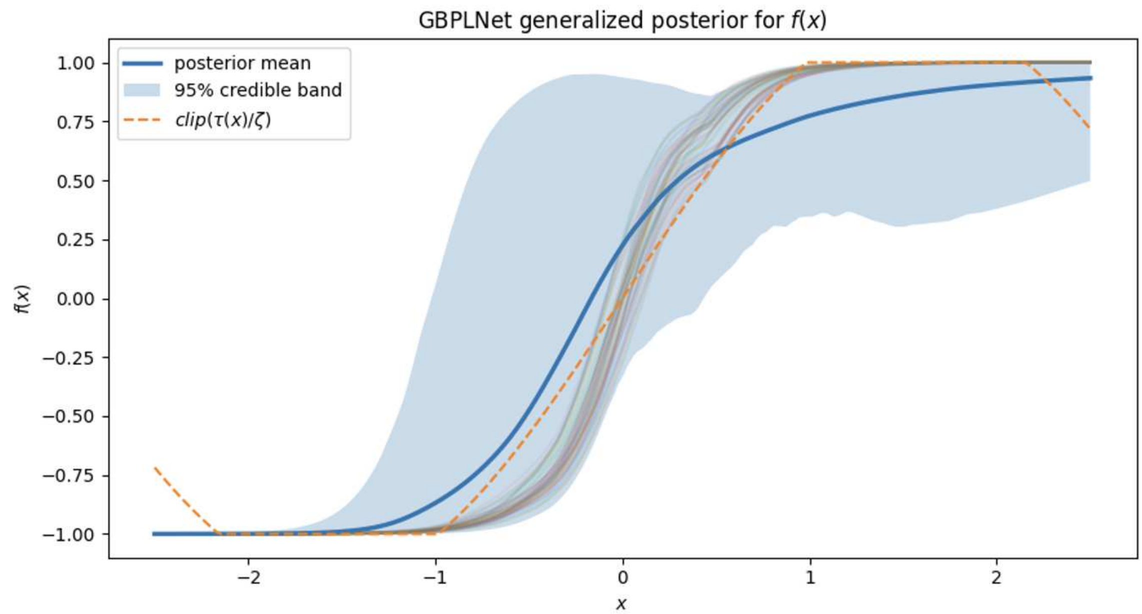

Figure 2: GBPLNet posterior visualization in the one-dimensional binary example. The plot shows posterior draws of the score function $f_w(x)$, the posterior mean, and a pointwise 95% credible band, along with the population target $\max(-1, \min(1, \tau(x)/\zeta))$.

H.4 POSTERIOR DISTRIBUTION OF WELFARE AND CREDIBLE INTERVALS

For each posterior draw $w^{(s)}$, we compute welfare on the test set using the deterministic policy $a_{w^{(s)}}(x) = \mathbb{1}[f_{w^{(s)}}(x) \geq 0]$. The welfare draw is

$$\widehat{V}^{(s)} = \frac{1}{n}\sum_{i=1}^{n}\Big(a_{w^{(s)}}(X_i)\,Y_i(1) + (1 - a_{w^{(s)}}(X_i))\,Y_i(0)\Big).$$

Figure 3 shows the empirical distribution of $\left\{\widehat{V}^{(s)}\right\}_{s=1}^{S}$. In this run, the posterior welfare mean is 0.4699, and the 95% credible interval given by the 0.025 and 0.975 quantiles is $[0.3327, 0.5021]$. These intervals summarize uncertainty under the generalized posterior. Their frequentist coverage depends on calibration choices for η and the prior, and they should be interpreted as loss-based posterior uncertainty summaries rather than as automatically calibrated confidence intervals.

H.4.1 How to use posterior uncertainty in practice

Posterior summaries can be used for several purposes. First, credible intervals for welfare provide an uncertainty report for the induced decision rule under the chosen loss scale. Second, posterior draws can be used to assess decision stability, for example, by inspecting posterior variability of the decision boundary through the distribution of $f_w(x_0)$ at selected covariate values. Figure 4 illustrates these distributions at five representative points. Third, if one prefers conservative action selection, one can compare policies by posterior lower quantiles of welfare computed on a validation set, which yields a robust selection criterion aligned with the welfare objective.

I ADDITIONAL EXPERIMENTS FOR LEARNING CURVES, WEAK OVERLAP, AND POSTERIOR UNCERTAINTY

This appendix reports additional experiments that complement the main simulation tables. The original full-feedback and counterfactual simulation tables are kept unchanged. The experiments here focus on three points raised in the reviews: learning curves under full feedback and logged feedback, behavior under weak overlap, and posterior uncertainty sum-

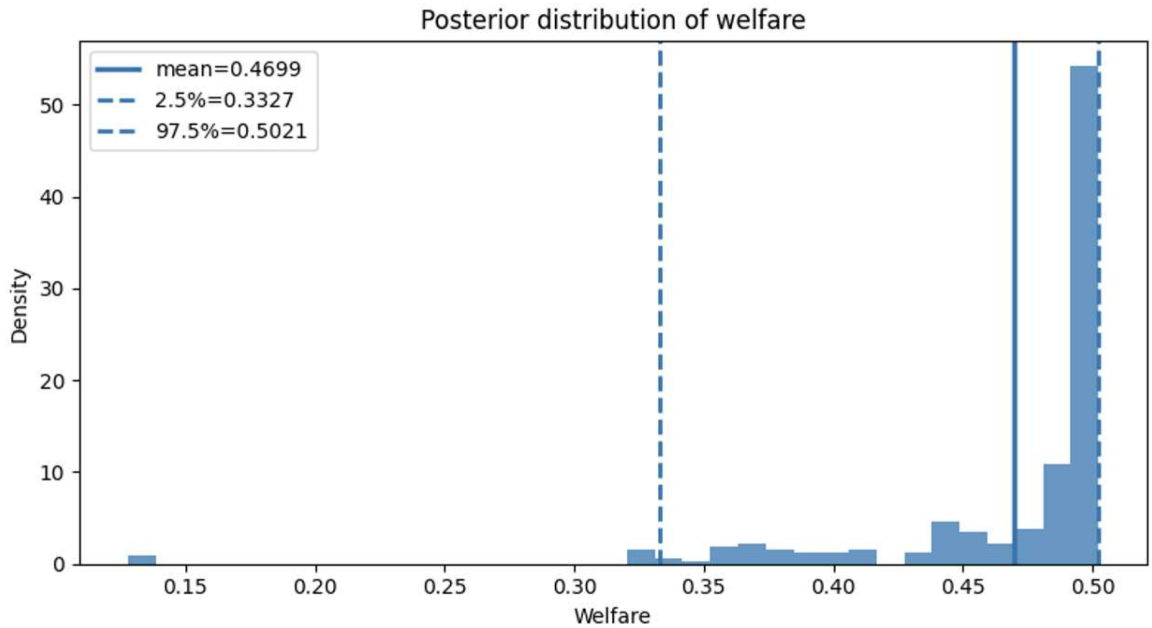

Figure 3: Posterior distribution of test welfare in the one-dimensional binary example. The histogram is computed from SGLD draws of w and welfare evaluation under the induced deterministic policy.

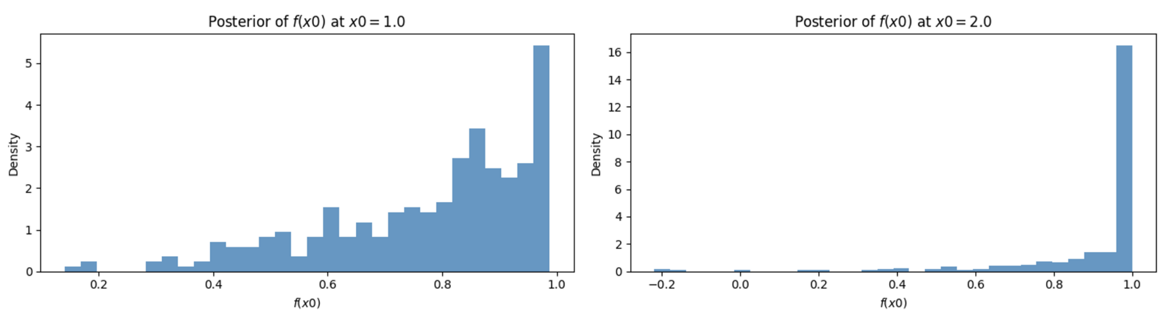

Figure 4: Posterior distributions of $f_w(x_0)$ at five fixed covariate values $x_0 \in \{-2, -1, 0, 1, 2\}$ in the one-dimensional binary example. These histograms visualize local decision uncertainty around the boundary $f_w(x_0) = 0$.

maries. All reported values are computed over independent replications. Values in parentheses are Monte Carlo standard errors.

I.1 LEARNING CURVES UNDER FULL FEEDBACK

We first report full-feedback learning curves for binary DGP2. For each observed sample size, we train the methods on the observed sample and evaluate the learned deterministic rule on an independent evaluation sample using the known conditional means. Figure 5 shows welfare and regret as functions of the observed sample size. Table 2 gives the corresponding numerical summaries.

I.2 LEARNING CURVES UNDER LOGGED FEEDBACK

We next report logged-feedback learning curves for binary DGP2. Only the outcome under the logged action is observed. GBPLNet-DR uses cross-fitted outcome regressions to form DR pseudo-outcomes, while GBPLNet-IPW uses IPW pseudo-outcomes. Figure 6 reports welfare and regret. Table 3 reports the numerical summaries.

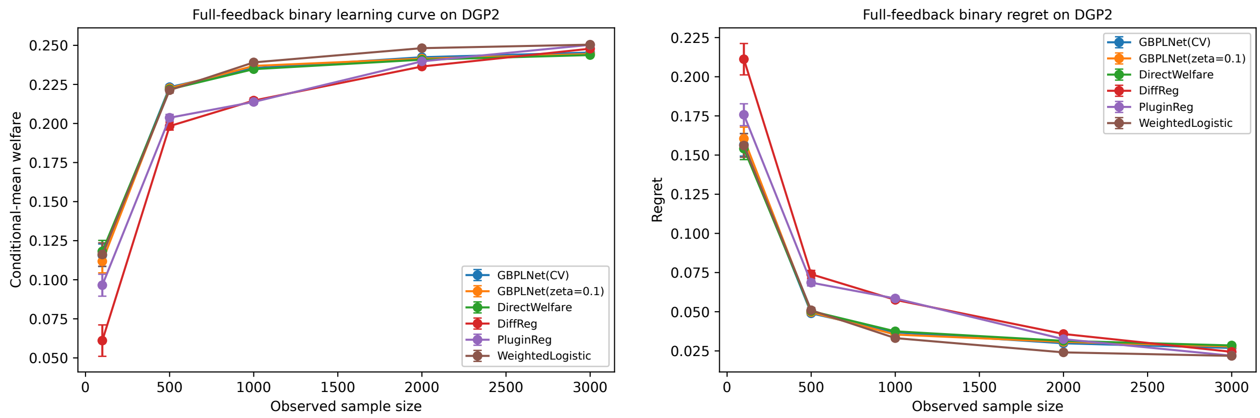

Figure 5: Full-feedback learning curves for binary DGP2 over 100 replications. The left panel reports welfare, and the right panel reports regret. Error bars show Monte Carlo standard errors.

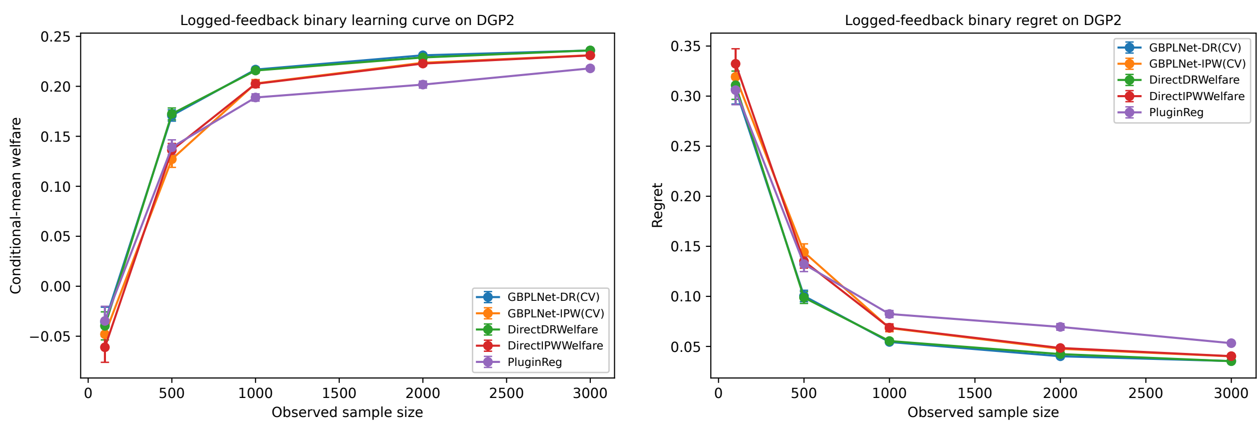

Figure 6: Logged-feedback learning curves for binary DGP2 over 50 replications. The left panel reports welfare, and the right panel reports regret. Error bars show Monte Carlo standard errors.

I.3 WEAK-OVERLAP DIAGNOSTICS

We examine the effect of propensity clipping under weak overlap. For each replication, we generate logged actions from a logistic policy with a stronger index and evaluate clipping levels $\epsilon_{\text{clip}} \in \{0.01, 0.05, 0.10\}$. Figure 7 shows welfare and the upper tail of the absolute pseudo-difference. Table 4 reports welfare, regret, the clipped fraction, the effective sample size, and pseudo-outcome tail summaries. Increasing the clipping level raises the effective sample size and reduces the tail of the pseudo-difference. For GBPLNet-DR (CV), welfare also increases slightly as clipping becomes more aggressive in this design.

I.4 REPEATED-SIMULATION EVALUATION OF WELFARE UNCERTAINTY

This subsection evaluates posterior welfare intervals in a one-dimensional binary design. For each replication, we fit GB-PLNet, construct a last-layer generalized Gauss–Newton approximation, draw 400 approximate posterior samples, and compute the welfare of each sampled policy on an independent evaluation sample. The resulting intervals are summaries from the loss-based posterior. Their frequentist coverage depends on the loss scale η, so we evaluate coverage by repeated simulation.

Table 5 reports calibration results for $\eta \in \{0.25, 0.5, 1, 2, 4, 8\}$ over 100 replications. The value $\eta = 4$ gives empirical coverage closest to 0.90 for the nominal 90% interval. Figure 8 shows the calibration curve. Table 6 then reports an independent evaluation over 200 new replications for $\eta = 1$ and $\eta = 4$. Calibration reduces interval width and brings the

Table 2: Full-feedback DGP2 learning-curve summary over 100 replications. Entries are means followed by Monte Carlo standard errors in parentheses.

n	Method	Welfare	Regret	Agreement
100	GBPLNet (CV)	0.1157 (0.0072)	0.1566 (0.0072)	0.7730 (0.0065)
	DirectWelfare	0.1182 (0.0070)	0.1541 (0.0070)	0.7752 (0.0063)
	DiffReg	0.0611 (0.0100)	0.2112 (0.0100)	0.7238 (0.0088)
	PluginReg	0.0965 (0.0070)	0.1757 (0.0070)	0.7551 (0.0063)
	WeightedLogistic	0.1162 (0.0075)	0.1561 (0.0075)	0.7729 (0.0069)
500	GBPLNet (CV)	0.2233 (0.0015)	0.0490 (0.0013)	0.8860 (0.0018)
	DirectWelfare	0.2218 (0.0013)	0.0504 (0.0011)	0.8834 (0.0016)
	DiffReg	0.1983 (0.0025)	0.0739 (0.0025)	0.8545 (0.0027)
	PluginReg	0.2037 (0.0021)	0.0686 (0.0021)	0.8615 (0.0023)
	WeightedLogistic	0.2213 (0.0018)	0.0509 (0.0017)	0.8826 (0.0024)
1000	GBPLNet (CV)	0.2354 (0.0009)	0.0368 (0.0007)	0.9047 (0.0012)
	DirectWelfare	0.2348 (0.0009)	0.0375 (0.0007)	0.9036 (0.0012)
	DiffReg	0.2147 (0.0017)	0.0576 (0.0015)	0.8762 (0.0019)
	PluginReg	0.2138 (0.0015)	0.0585 (0.0013)	0.8756 (0.0015)
	WeightedLogistic	0.2391 (0.0011)	0.0332 (0.0009)	0.9109 (0.0016)
2000	GBPLNet (CV)	0.2424 (0.0008)	0.0299 (0.0005)	0.9182 (0.0010)
	DirectWelfare	0.2408 (0.0008)	0.0314 (0.0005)	0.9148 (0.0010)
	DiffReg	0.2365 (0.0010)	0.0358 (0.0006)	0.9053 (0.0009)
	PluginReg	0.2398 (0.0010)	0.0325 (0.0006)	0.9108 (0.0009)
	WeightedLogistic	0.2482 (0.0008)	0.0241 (0.0004)	0.9315 (0.0010)
3000	GBPLNet (CV)	0.2454 (0.0008)	0.0269 (0.0004)	0.9248 (0.0009)
	DirectWelfare	0.2438 (0.0008)	0.0285 (0.0003)	0.9207 (0.0007)
	DiffReg	0.2479 (0.0008)	0.0244 (0.0003)	0.9231 (0.0006)
	PluginReg	0.2503 (0.0007)	0.0219 (0.0003)	0.9272 (0.0005)
	WeightedLogistic	0.2504 (0.0007)	0.0218 (0.0003)	0.9378 (0.0008)

90% and 95% intervals close to their nominal coverage levels in this design.

I.5 POSTERIOR ACTION DISAGREEMENT

We next examine whether posterior uncertainty is large in regions where the learned action is unstable. For each evaluation point, define $p_{\mathcal{D}}(x) = \mathbb{P}_{w \sim \Pi_\eta(\cdot|\mathcal{D})}(f_w(x) \geq 0)$ and $u_{\mathcal{D}}(x) = 2\min\{p_{\mathcal{D}}(x), 1 - p_{\mathcal{D}}(x)\}$. We approximate $p_{\mathcal{D}}(x)$ using the same last-layer generalized Gauss–Newton draws as above. The score $u_{\mathcal{D}}(x)$ is close to zero when posterior draws mostly agree on the action and close to one when posterior draws are split.

Let $\widehat{d}_{\mathcal{D}}(x) = \mathbb{1}[f_{\widehat{w}_{\text{MAP}}}(x) \geq 0]$ and $d^\star(x) = \mathbb{1}[\tau(x) \geq 0]$. We define the conditional margin by $m(x) = |\tau(x)|$ and the regret contribution by $r_{\mathcal{D}}(x) = m(x)\,\mathbb{1}\left[\widehat{d}_{\mathcal{D}}(x) \neq d^\star(x)\right]$. We divide evaluation points into five groups according to $u_{\mathcal{D}}(x) \in [0, 0.2)$, $u_{\mathcal{D}}(x) \in [0.2, 0.4)$, $u_{\mathcal{D}}(x) \in [0.4, 0.6)$, $u_{\mathcal{D}}(x) \in [0.6, 0.8)$, and $u_{\mathcal{D}}(x) \in [0.8, 1]$. Table 7 reports the results. Higher posterior disagreement is associated with lower agreement with the Bayes-optimal action and a larger regret contribution. Figure 10 shows the same pattern graphically.

I.6 POSTERIOR-THRESHOLD RULES

The MAP rule is a point decision and does not directly use the width of the welfare interval. The posterior can also define uncertainty-aware rules. For $\alpha \in \{0.05, 0.10, 0.20, 0.30\}$, we consider the threshold rule that selects action 1 when $p_{\mathcal{D}}(x) > 1 - \alpha$, selects action 0 when $p_{\mathcal{D}}(x) < \alpha$, and otherwise uses action 0 as the default. The uncertain-region rate is the fraction of evaluation points satisfying $\alpha \leq p_{\mathcal{D}}(x) \leq 1 - \alpha$, and the action-change rate is the fraction for which the

Table 3: Logged-feedback DGP2 learning-curve summary over 50 replications. Entries are means followed by Monte Carlo standard errors in parentheses.

n	Method	Welfare	Regret	Agreement
100	GBPLNet-DR (CV)	-0.0356 (0.0146)	0.3068 (0.0147)	0.6405 (0.0126)
	GBPLNet-IPW (CV)	-0.0480 (0.0140)	0.3192 (0.0140)	0.6299 (0.0119)
	DirectDRWelfare	-0.0397 (0.0140)	0.3109 (0.0141)	0.6367 (0.0120)
	DirectIPWWelfare	-0.0611 (0.0151)	0.3323 (0.0150)	0.6185 (0.0129)
	PluginReg	-0.0347 (0.0147)	0.3060 (0.0146)	0.6411 (0.0124)
500	GBPLNet-DR (CV)	0.1708 (0.0056)	0.1004 (0.0057)	0.8267 (0.0057)
	GBPLNet-IPW (CV)	0.1271 (0.0081)	0.1442 (0.0083)	0.7851 (0.0077)
	DirectDRWelfare	0.1724 (0.0059)	0.0988 (0.0059)	0.8284 (0.0059)
	DirectIPWWelfare	0.1361 (0.0068)	0.1351 (0.0069)	0.7931 (0.0066)
	PluginReg	0.1390 (0.0075)	0.1322 (0.0075)	0.7950 (0.0069)
1000	GBPLNet-DR (CV)	0.2168 (0.0022)	0.0544 (0.0017)	0.8777 (0.0023)
	GBPLNet-IPW (CV)	0.2029 (0.0039)	0.0684 (0.0037)	0.8624 (0.0042)
	DirectDRWelfare	0.2158 (0.0024)	0.0554 (0.0022)	0.8773 (0.0029)
	DirectIPWWelfare	0.2025 (0.0037)	0.0687 (0.0038)	0.8621 (0.0044)
	PluginReg	0.1888 (0.0036)	0.0824 (0.0035)	0.8458 (0.0038)
2000	GBPLNet-DR (CV)	0.2311 (0.0017)	0.0402 (0.0013)	0.8993 (0.0021)
	GBPLNet-IPW (CV)	0.2236 (0.0027)	0.0476 (0.0023)	0.8890 (0.0032)
	DirectDRWelfare	0.2289 (0.0015)	0.0423 (0.0010)	0.8950 (0.0017)
	DirectIPWWelfare	0.2228 (0.0022)	0.0484 (0.0017)	0.8872 (0.0025)
	PluginReg	0.2017 (0.0035)	0.0695 (0.0033)	0.8628 (0.0037)
3000	GBPLNet-DR (CV)	0.2359 (0.0017)	0.0354 (0.0012)	0.9079 (0.0021)
	GBPLNet-IPW (CV)	0.2311 (0.0016)	0.0401 (0.0014)	0.8999 (0.0021)
	DirectDRWelfare	0.2360 (0.0015)	0.0353 (0.0009)	0.9073 (0.0016)
	DirectIPWWelfare	0.2309 (0.0018)	0.0403 (0.0015)	0.8996 (0.0023)
	PluginReg	0.2180 (0.0021)	0.0533 (0.0017)	0.8824 (0.0020)

threshold rule differs from the MAP rule.

Table 8 reports the results. In this design, the threshold rule changes a small fraction of actions and has a small welfare loss relative to the MAP rule. Figure 11 plots the welfare difference against the uncertain-region rate.

J EXPERIMENTAL DETAILS

J.1 COMMON PROTOCOL

This subsection summarizes the experimental protocol used in Section 9. In all synthetic experiments, we observe the full feedback vector for each unit, and welfare is computed on a held-out test set using the realized potential outcomes. We repeat each experiment 100 times with independent random seeds. For each trial, we split the data into training, validation, and test sets with proportions 0.6, 0.2, and 0.2. We report the mean and variance of the test welfare across trials, together with boxplots. For GBPL and GBPLNet, we compare four values of ζ, namely $\zeta \in \{1, 0.1, 0.01\}$ and a validation-selected value among a fixed grid.

J.2 SYNTHETIC EXPERIMENTS WITH BINARY ACTIONS

In the binary action experiments, we generate i.i.d. covariates $X \in \mathbb{R}^d$ with $d = 10$ from a standard normal distribution. For each draw, we generate two potential outcomes $Y(0)$ and $Y(1)$ with additive Gaussian noise. We consider three data generating processes.

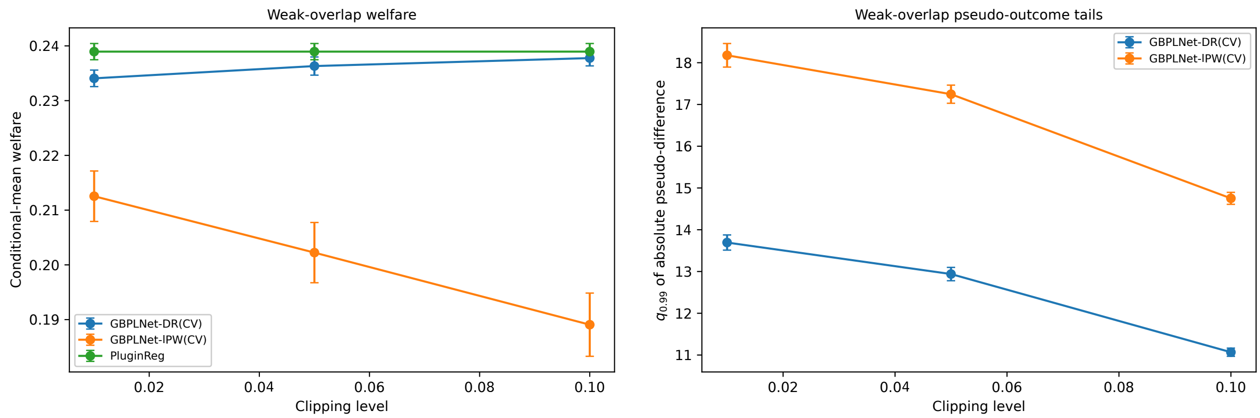

Figure 7: Weak-overlap diagnostics over 50 replications. The left panel reports welfare by clipping level. The right panel reports the 0.99 quantile of the absolute pseudo-difference. Error bars show Monte Carlo standard errors.

Table 4: Weak-overlap diagnostics over 50 replications. Welfare and regret are means followed by Monte Carlo standard errors in parentheses. The clipped fraction, ESS, $q_{0.99}$, and maximum are computed from the training portion of the logged sample.

Clip	Method	Welfare	Regret	Clipped fraction	ESS	$q_{0.99}\left(\left\|\widetilde{U}\right\|\right)$	$\max_i \left\|\widetilde{U}_i\right\|$
0.01	GBPLNet-DR (CV)	0.2341 (0.0015)	0.0371 (0.0010)	0.0684	1090.1	13.6938	179.0290
	GBPLNet-IPW (CV)	0.2125 (0.0046)	0.0587 (0.0042)	0.0684	1090.1	18.1773	196.5811
	PluginReg	0.2390 (0.0015)	0.0323 (0.0008)	0.0684	1090.1	–	–
0.05	GBPLNet-DR (CV)	0.2363 (0.0017)	0.0349 (0.0011)	0.2433	2071.9	12.9393	57.6668
	GBPLNet-IPW (CV)	0.2022 (0.0055)	0.0690 (0.0052)	0.2433	2071.9	17.2457	66.3725
	PluginReg	0.2390 (0.0015)	0.0323 (0.0008)	0.2433	2071.9	–	–
0.10	GBPLNet-DR (CV)	0.2378 (0.0014)	0.0334 (0.0008)	0.3845	2760.7	11.0675	32.1321
	GBPLNet-IPW (CV)	0.1891 (0.0058)	0.0821 (0.0054)	0.3845	2760.7	14.7522	39.0025
	PluginReg	0.2390 (0.0015)	0.0323 (0.0008)	0.3845	2760.7	–	–

In DGP1, the baseline outcome mean is $\gamma_0(x) = x_1 + 0.5x_2^2 - 0.25x_3x_4$ and the conditional treatment effect is $\tau(x) = 2\tanh\left(x^\top w/\sqrt{d}\right)$ for a fixed unit-norm vector w. We set $Y(0) = \gamma_0(X) + \varepsilon_0$ and $Y(1) = \gamma_0(X) + \tau(X) + \varepsilon_1$ with independent $\varepsilon_0, \varepsilon_1 \sim \mathcal{N}(0, 1)$.

In DGP2, the baseline outcome mean is $\gamma_0(x) = 0.5\sin(x_1) + 0.3x_2 - 0.2x_3^2$ and the treatment effect is $\tau(x) = 1.5\sin(x_1 + x_2)$, and we generate outcomes as above with independent standard normal noise.

In DGP3, the baseline outcome mean is $\gamma_0(x) = 0.2x_1 - 0.1x_2 + 0.1\tanh(x_3)$ and the treatment effect is $\tau(x) = 2.5\left(\mathbb{1}[x_1 > 0] - 0.5 + 0.2x_2\right)$, and we generate outcomes as above with independent standard normal noise.

For evaluation, a fitted score $f(x) \in (-1, 1)$ induces a deterministic policy that selects action 1 if $f(x) \geq 0$ and selects action 0 otherwise. The oracle policy selects the action with the larger realized potential outcome.

J.3 SYNTHETIC EXPERIMENTS WITH MULTIPLE ACTIONS

In the multi-action experiments, we set $K = 5$ and generate i.i.d. covariates $X \in \mathbb{R}^d$ with $d = 10$ from a standard normal distribution. For each unit, we generate the full feedback vector $(Y(1), \ldots, Y(K))$ with additive Gaussian noise. We again consider three data generating processes.

In DGP1, the baseline component is $b(x) = x_1 + 0.5x_2^2 - 0.25x_3x_4$, and for each action $a \in \{1, \ldots, K\}$ we generate a random unit-norm vector w_a and define $\gamma_a(x) = b(x) + 1.5\sin\left(x^\top w_a/\sqrt{d} + 0.3a\right)$. We set $Y(a) = \gamma_a(X) + \varepsilon_a$ with

Table 5: Welfare-interval calibration over 100 replications. Each entry is empirical coverage followed by mean interval width in parentheses.

η	50%	80%	90%	95%
0.25	0.93 (0.0479)	1.00 (0.1114)	1.00 (0.1670)	1.00 (0.2219)
0.5	0.78 (0.0376)	0.99 (0.0822)	0.99 (0.1195)	1.00 (0.1566)
1	0.69 (0.0274)	0.95 (0.0604)	0.99 (0.0869)	0.99 (0.1152)
2	0.63 (0.0195)	0.88 (0.0426)	0.97 (0.0615)	0.99 (0.0830)
4	0.54 (0.0135)	0.80 (0.0289)	0.90 (0.0413)	0.93 (0.0545)
8	0.47 (0.0093)	0.72 (0.0192)	0.80 (0.0265)	0.84 (0.0342)

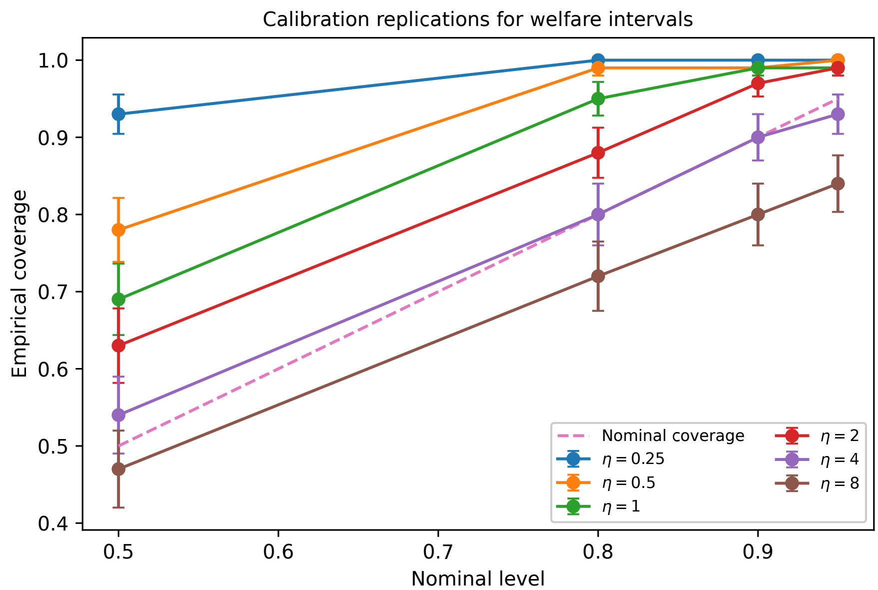

Figure 8: Empirical welfare-interval coverage on the 100 calibration replications.

independent $\varepsilon_a \sim \mathcal{N}(0,1)$.

In DGP2, the baseline component is $b(x) = 0.5\sin(x_1) + 0.3x_2 - 0.2x_3^2$, and $\gamma_a(x) = b(x) + 2\tanh\left(x^\top w_a/\sqrt{d} - 0.2a\right)$.

In DGP3, the baseline component is $b(x) = 0.2x_1 - 0.1x_2 + 0.1\tanh(x_3)$, and $\gamma_a(x) = b(x) + 1.0\mathbb{1}\left[x^\top w_a/\sqrt{d} > 0\right] + 0.1a$.

A fitted model outputs a probability vector $\delta(x) \in \Delta_K$, and we evaluate the corresponding stochastic policy whose welfare is given by $\mathbb{E}\left[\sum_{a\in\{1,\ldots,K\}} \delta_a(X)\, Y(a)\right]$, which is estimated by the sample average on the test set.

J.4 UCI EXPERIMENTS

The UCI/OpenML experiments in Section 9 use real covariates and a real-valued response y obtained from OpenML [Vanschoren et al., 2013] and the UCI repository [Dua and Graff, 2019]. Because these benchmark datasets provide only a single response, we construct synthetic full feedback vectors by defining action-dependent shifts as smooth functions of x and adding them to a standardized version of y. Concretely, we construct $Y(a)$ by adding an action-specific term of the form $\tanh\left(x^\top w_a/\sqrt{d} + c_a\right)$ to the standardized response, and we then evaluate welfare using the resulting full feedback vector. This design keeps the covariate distribution and marginal response distribution realistic, while creating heterogeneous action effects that allow policy learning to be evaluated under full-outcome feedback.

Table 6: Independent welfare-interval evaluation over 200 replications. Each entry is empirical coverage followed by mean interval width in parentheses.

η	50%	80%	90%	95%
1	0.750 (0.0283)	0.945 (0.0624)	0.980 (0.0913)	0.990 (0.1231)
4	0.635 (0.0139)	0.855 (0.0300)	0.910 (0.0433)	0.950 (0.0575)

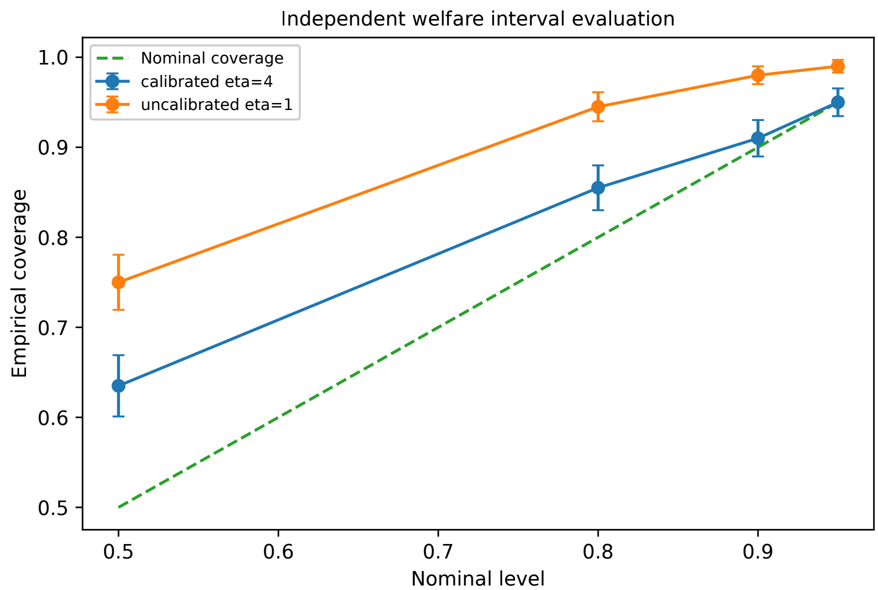

Figure 9: Independent welfare-interval evaluation over 200 replications. The dashed line is nominal coverage.

K ADDITIONAL EXPERIMENTAL RESULTS WITH FULL-FEEDBACK SETTINGS

K.1 EXPERIMENTS WITH MULTIPLE ACTIONS

We consider $K = 5$ actions and the baseline-free symmetric full-vector surrogate in Section 5. Each trial uses $n = 6000$ observations and $d = 10$ features. We compare DirectWelfare, a direct optimization baseline that maximizes empirical welfare in the policy class, and GBPLNet, the proposed General Bayes method based on the full-vector squared-loss surrogate. For GBPLNet, we again compare $\zeta \in \{1, 0.1, 0.01\}$ and a validation choice from $\{1, 0.1, 0.01, 0.001\}$. Table 9 reports welfare and regret, and Figure 12 shows welfare boxplots.

Table 9 indicates that GBPLNet is competitive with DirectWelfare across the three scenarios. The results also show that the choice of ζ can matter for welfare. In particular, in DGP3, $\zeta = 1.0$ yields noticeably lower welfare, while selection by validation welfare avoids this deterioration. Across the 100 DGP3 trials, the validation rule selects $\zeta = 0.001$, 0.01, and 0.1 in 5, 45, and 50 trials, respectively, and never selects $\zeta = 1.0$.

K.2 UCI AND OPENML EXPERIMENTS

We also evaluate the binary action setting on two regression datasets, yacht and energy_efficiency, accessed via OpenML [Vanschoren et al., 2013], and originally hosted in the UCI Machine Learning Repository [Dua and Graff, 2019]. Since these datasets provide a single response variable, we construct full potential outcomes by adding an action dependent effect to the original response, so that each observation contains $(Y(0), Y(1))$. This construction yields a controlled full feedback benchmark that preserves the feature distribution of each dataset.

We run 100 trials using random splits, and we compare the same methods as in the synthetic binary experiments. Table 10 reports welfare and regret, and Figure 13 shows welfare boxplots. In these two datasets, all methods achieve similar welfare,

Table 7: Posterior action disagreement over 200 replications. Points is the pooled number of evaluation points in each group. The remaining entries are averages of replication-level group means followed by Monte Carlo standard errors in parentheses.

Group	Points	Disagreement	Agreement	Margin	Regret
1	391154	0.0023 (0.0001)	0.9955 (0.0005)	0.8822 (0.0002)	0.0002 (0.0000)
2	3026	0.2923 (0.0005)	0.7176 (0.0173)	0.0863 (0.0032)	0.0165 (0.0017)
3	2210	0.4958 (0.0006)	0.6568 (0.0153)	0.0802 (0.0037)	0.0219 (0.0018)
4	1859	0.6972 (0.0007)	0.5799 (0.0122)	0.0762 (0.0039)	0.0278 (0.0019)
5	1751	0.8992 (0.0006)	0.5225 (0.0103)	0.0748 (0.0040)	0.0348 (0.0023)

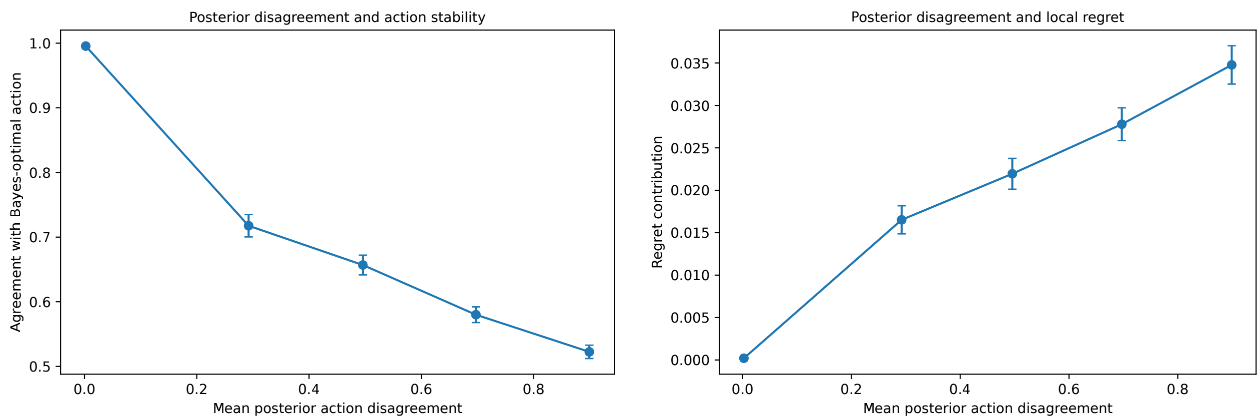

Figure 10: Posterior action disagreement over 200 replications. The left panel reports agreement with the Bayes-optimal action. The right panel reports the regret contribution. Error bars show Monte Carlo standard errors.

with extremely small regret relative to the oracle benchmark. DiffReg is slightly better on average, and WeightedLogistic is slightly worse.

L ADDITIONAL EXPERIMENTAL RESULTS WITH COUNTERFACTUAL SETTINGS

This appendix reports additional empirical results for the missing outcome setting in Section 7. In these experiments, the learner observes only the logged outcome $Y_i = Y_i(A_i)$ and the action A_i, and the full outcome vector is used only for evaluation on a held out test set.

L.1 COMMON PROTOCOL

We follow the same training protocol as in Appendix J. Each trial splits the data into training, validation, and test sets with proportions 0.6, 0.2, and 0.2. For each candidate model, we train on the training set and select the epoch by early stopping on the validation loss. For the surrogate tuning parameter, we compare $\zeta \in \{1, 0.1, 0.01\}$ and a validation-selected choice from the grid $\{1, 0.1, 0.01, 0.001\}$. In the tables, the label (CV) indicates the validation-selected ζ. Because the full outcome vector is not observed in this setting, we select ζ by the IPW or DR pseudo-welfare on the validation set, matching the pseudo-outcome construction used for training.

L.2 LOGGED DATA CONSTRUCTION AND PSEUDO-OUTCOMES

The synthetic experiments generate potential outcomes as in Appendix J and then convert them into logged data by drawing actions from a stochastic logging policy. For binary actions, we generate a propensity score $e(x) = \mathbb{P}(A = 1 \mid X = x)$ from a logistic model with a random linear index, and we clip $e(x)$ to satisfy overlap. For K actions, we generate propensities $\{e_a(x)\}_{a=1}^{K}$ from a softmax model with random linear logits, and we sample A from the resulting multinomial

Table 8: Posterior-threshold rules over 200 replications. Welfare and Difference are means followed by Monte Carlo standard errors in parentheses. The fifth percentile is computed across replications.

α	Welfare	5th pct.	MAP	Difference	Uncertain	Change
0.05	0.4309 (0.0001)	0.4278	0.4315	-0.0006 (0.0001)	0.0285	0.0144
0.10	0.4311 (0.0001)	0.4287	0.4315	-0.0004 (0.0001)	0.0222	0.0112
0.20	0.4313 (0.0001)	0.4293	0.4315	-0.0002 (0.0001)	0.0146	0.0074
0.30	0.4314 (0.0001)	0.4293	0.4315	-0.0001 (0.0000)	0.0090	0.0046

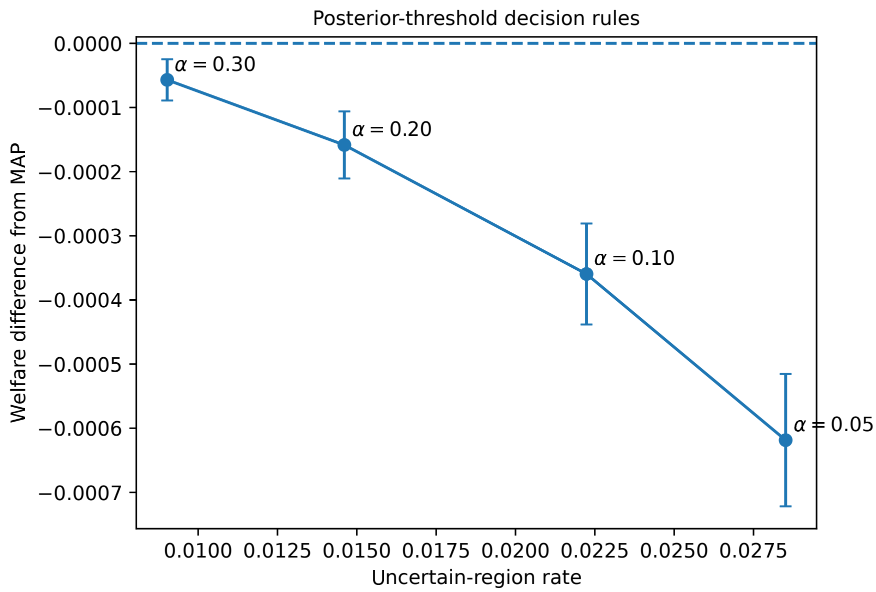

Figure 11: Posterior-threshold rules over 200 replications. The vertical axis reports the welfare difference from the MAP rule.

distribution.

Given logged data $\{(X_i, A_i, Y_i)\}_{i=1}^n$, we compute pseudo-outcomes by IPW estimator in (16) or by the DR construction in (17). In the binary action case, we form a pseudo-outcome difference by $\widetilde{U}_i = \widetilde{Y}_i(1) - \widetilde{Y}_i(0)$, and we train a bounded score model by minimizing the squared-loss surrogate in (6) with $Y_i(1) - Y_i(0)$ replaced by $\widetilde{U}_i$. In the multi-action case, we use the full-vector empirical losses $\widehat{L}_n^{\mathrm{IPW}}(\theta)$ and its DR analog.

For DR pseudo-outcomes, we estimate the nuisance outcome regression $\widehat{\gamma}_a(x)$ by a neural network trained with a masked squared loss on observed outcomes. For propensity scores, we use the true propensity in the synthetic experiments to isolate the effect of the surrogate and the pseudo-outcome construction. In the UCI and OpenML experiments, the propensity is estimated by multinomial or binary logistic regression on the training set.

L.3 SYNTHETIC EXPERIMENTS WITH BINARY ACTIONS

We use the three synthetic DGPs described in Appendix J and generate logged data with $n = 6000$ and $d = 10$. We repeat the experiment for 50 trials. Table 11 reports welfare and regret.

The comparison methods are as follows. GBPLNet-DR and GBPLNet-IPW apply the proposed squared-loss surrogate with DR and IPW pseudo-outcome differences. PluginReg fits a nuisance outcome regression model and chooses the action with the larger predicted outcome. Since the outcome difference is not observed, DiffReg is not applicable in this setting.

Across the three DGPs, DR based GBPLNet is typically more stable than IPW based GBPLNet. The sensitivity to ζ remains scenario dependent, and the validation-selected choice performs competitively in all three DGPs in this run.

Table 9: Synthetic $K = 5$ results, 100 trials. Columns report welfare mean, welfare variance, welfare standard error, regret mean, and regret standard error across trials.

dataset	method	welfare_mean	welfare_var	welfare_se	regret_mean	regret_se
DGP1	DirectWelfare	1.3366	0.0117	0.0108	0.4972	0.0066
	GBPLNet ($\zeta = 0.01$)	1.3404	0.0115	0.0107	0.4934	0.0063
	GBPLNet ($\zeta = 0.1$)	1.3432	0.0105	0.0102	0.4906	0.0056
	GBPLNet ($\zeta = 1.0$)	1.3417	0.0095	0.0097	0.4921	0.0050
	GBPLNet (CV)	1.3588	0.0098	0.0099	0.4751	0.0051
DGP2	DirectWelfare	1.4239	0.0096	0.0098	0.4850	0.0052
	GBPLNet ($\zeta = 0.01$)	1.4245	0.0090	0.0095	0.4844	0.0049
	GBPLNet ($\zeta = 0.1$)	1.4226	0.0101	0.0100	0.4864	0.0057
	GBPLNet ($\zeta = 1.0$)	1.4154	0.0103	0.0102	0.4936	0.0048
	GBPLNet (CV)	1.4379	0.0095	0.0098	0.4710	0.0046
DGP3	DirectWelfare	0.8385	0.0031	0.0056	0.5136	0.0043
	GBPLNet ($\zeta = 0.01$)	0.8389	0.0031	0.0055	0.5131	0.0043
	GBPLNet ($\zeta = 0.1$)	0.8403	0.0029	0.0054	0.5117	0.0041
	GBPLNet ($\zeta = 1.0$)	0.7466	0.0023	0.0048	0.6054	0.0037
	GBPLNet (CV)	0.8413	0.0028	0.0053	0.5107	0.0041

L.4 SYNTHETIC EXPERIMENTS WITH MULTIPLE ACTIONS

We next consider $K = 5$ actions and use the three multi-action DGPs described in Appendix J. We generate logged data with $n = 8000$ and $d = 10$, and we repeat the experiment for 50 trials. Table 12 reports welfare and regret.

In this setting, we use the baseline-free full-vector surrogate in Section 5 and the missing-outcome empirical losses in Section 7. We implement the policy class by a softmax network that outputs $\pi_\theta(x) \in \Delta_K$, and we refer to the resulting method as GBPL-full. The comparison method PluginRegK fits a nuisance outcome regression model and chooses the action with the largest predicted outcome.

L.5 UCI AND OPENML EXPERIMENTS

We also report counterfactual experiments on the regression datasets yacht and energy_efficiency, accessed via OpenML. Since these datasets provide a single response variable, we follow the same semi synthetic construction as in Appendix K, we generate potential outcomes by adding an action dependent effect to the original response. We then generate logged data by sampling actions from a stochastic logging policy and observing only $Y(A)$. We repeat the experiment for 30 trials with random splits. Table 13 reports welfare and regret.

Remark. *The counterfactual experiments illustrate how the framework in Section 7 can be implemented with standard nuisance estimation and validation. In applications with logged outcomes, one can proceed as follows. First, choose a policy class, either a bounded score model $f_\theta(x)$ for binary actions or a simplex-valued model $\pi_\theta(x)$ for multiple actions. Second, estimate propensities $\widehat{e}_a(x)$ and, for DR, outcome regression functions $\widehat{\gamma}_a(x)$ using flexible supervised learners. Third, construct IPW or DR pseudo-outcomes and minimize the corresponding empirical loss. Finally, select ζ by validation, and report welfare using an off policy evaluation estimator that matches the pseudo-outcome construction. In settings where posterior uncertainty is desired, the same empirical losses can be used in generalized posterior sampling by the approximation methods described in Section 6.*

Table 10: UCI and OpenML results, binary actions, 100 trials. Columns report welfare mean, welfare variance, welfare standard error, regret mean, and regret standard error across trials.

dataset	method	welfare_mean	welfare_var	welfare_se	regret_mean	regret_se
energy_efficiency	DiffReg	0.3569	0.0295	0.0172	0.0001	0.0000
	GBPLNet ($\zeta = 0.01$)	0.3565	0.0295	0.0172	0.0004	0.0001
	GBPLNet ($\zeta = 0.1$)	0.3568	0.0295	0.0172	0.0002	0.0001
	GBPLNet ($\zeta = 1.0$)	0.3568	0.0295	0.0172	0.0002	0.0000
	GBPLNet (CV)	0.3568	0.0295	0.0172	0.0002	0.0000
	PluginReg	0.3567	0.0296	0.0172	0.0002	0.0000
	WeightedLogistic	0.3553	0.0296	0.0172	0.0016	0.0002
yacht	DiffReg	0.3465	0.0234	0.0153	0.0001	0.0000
	GBPLNet ($\zeta = 0.01$)	0.3458	0.0234	0.0153	0.0008	0.0002
	GBPLNet ($\zeta = 0.1$)	0.3463	0.0234	0.0153	0.0004	0.0001
	GBPLNet ($\zeta = 1.0$)	0.3464	0.0234	0.0153	0.0002	0.0001
	GBPLNet (CV)	0.3464	0.0234	0.0153	0.0003	0.0001
	PluginReg	0.3463	0.0234	0.0153	0.0003	0.0000
	WeightedLogistic	0.3449	0.0233	0.0153	0.0017	0.0002

Table 11: Counterfactual synthetic binary results, 50 trials. Columns report welfare mean, welfare variance, welfare standard error, regret mean, and regret standard error across trials. GBPLNet-DR and GBPLNet-IPW use DR and IPW pseudo-outcome differences, respectively.

dgp	method	welfare_mean	welfare_var	welfare_se	regret_mean	regret_se
DGP1	GBPLNet-DR ($\zeta = 0.01$)	1.0184	0.0086	0.0131	0.2336	0.0055
DGP1	GBPLNet-DR ($\zeta = 0.1$)	1.0161	0.0082	0.0128	0.2359	0.0051
DGP1	GBPLNet-DR ($\zeta = 1.0$)	1.0212	0.0088	0.0133	0.2309	0.0054
DGP1	GBPLNet-DR (CV)	1.0211	0.0081	0.0127	0.2309	0.0050
DGP1	GBPLNet-IPW ($\zeta = 0.01$)	1.0041	0.0085	0.0130	0.2479	0.0054
DGP1	GBPLNet-IPW ($\zeta = 0.1$)	1.0002	0.0082	0.0128	0.2518	0.0054
DGP1	GBPLNet-IPW ($\zeta = 1.0$)	1.0068	0.0082	0.0128	0.2452	0.0052
DGP1	GBPLNet-IPW (CV)	1.0075	0.0081	0.0127	0.2445	0.0052
DGP1	PluginReg	1.0198	0.0077	0.0088	0.2322	0.0033
DGP2	GBPLNet-DR ($\zeta = 0.01$)	0.7096	0.0024	0.0069	0.3296	0.0052
DGP2	GBPLNet-DR ($\zeta = 0.1$)	0.7063	0.0025	0.0071	0.3329	0.0053
DGP2	GBPLNet-DR ($\zeta = 1.0$)	0.6970	0.0025	0.0071	0.3422	0.0055
DGP2	GBPLNet-DR (CV)	0.6984	0.0026	0.0072	0.3408	0.0051
DGP2	GBPLNet-IPW ($\zeta = 0.01$)	0.6967	0.0022	0.0066	0.3425	0.0055
DGP2	GBPLNet-IPW ($\zeta = 0.1$)	0.6896	0.0026	0.0072	0.3496	0.0051
DGP2	GBPLNet-IPW ($\zeta = 1.0$)	0.6778	0.0027	0.0073	0.3613	0.0058
DGP2	GBPLNet-IPW (CV)	0.6780	0.0023	0.0068	0.3612	0.0052
DGP2	PluginReg	0.6720	0.0031	0.0056	0.3672	0.0040
DGP3	GBPLNet-DR ($\zeta = 0.01$)	0.4744	0.0052	0.0102	0.3057	0.0043
DGP3	GBPLNet-DR ($\zeta = 0.1$)	0.4757	0.0055	0.0105	0.3044	0.0045
DGP3	GBPLNet-DR ($\zeta = 1.0$)	0.4744	0.0053	0.0103	0.3056	0.0042
DGP3	GBPLNet-DR (CV)	0.4764	0.0053	0.0103	0.3036	0.0041
DGP3	GBPLNet-IPW ($\zeta = 0.01$)	0.4729	0.0058	0.0107	0.3072	0.0045
DGP3	GBPLNet-IPW ($\zeta = 0.1$)	0.4695	0.0057	0.0107	0.3106	0.0046
DGP3	GBPLNet-IPW ($\zeta = 1.0$)	0.4736	0.0058	0.0108	0.3065	0.0045
DGP3	GBPLNet-IPW (CV)	0.4744	0.0053	0.0103	0.3056	0.0040
DGP3	PluginReg	0.4749	0.0049	0.0070	0.3051	0.0026

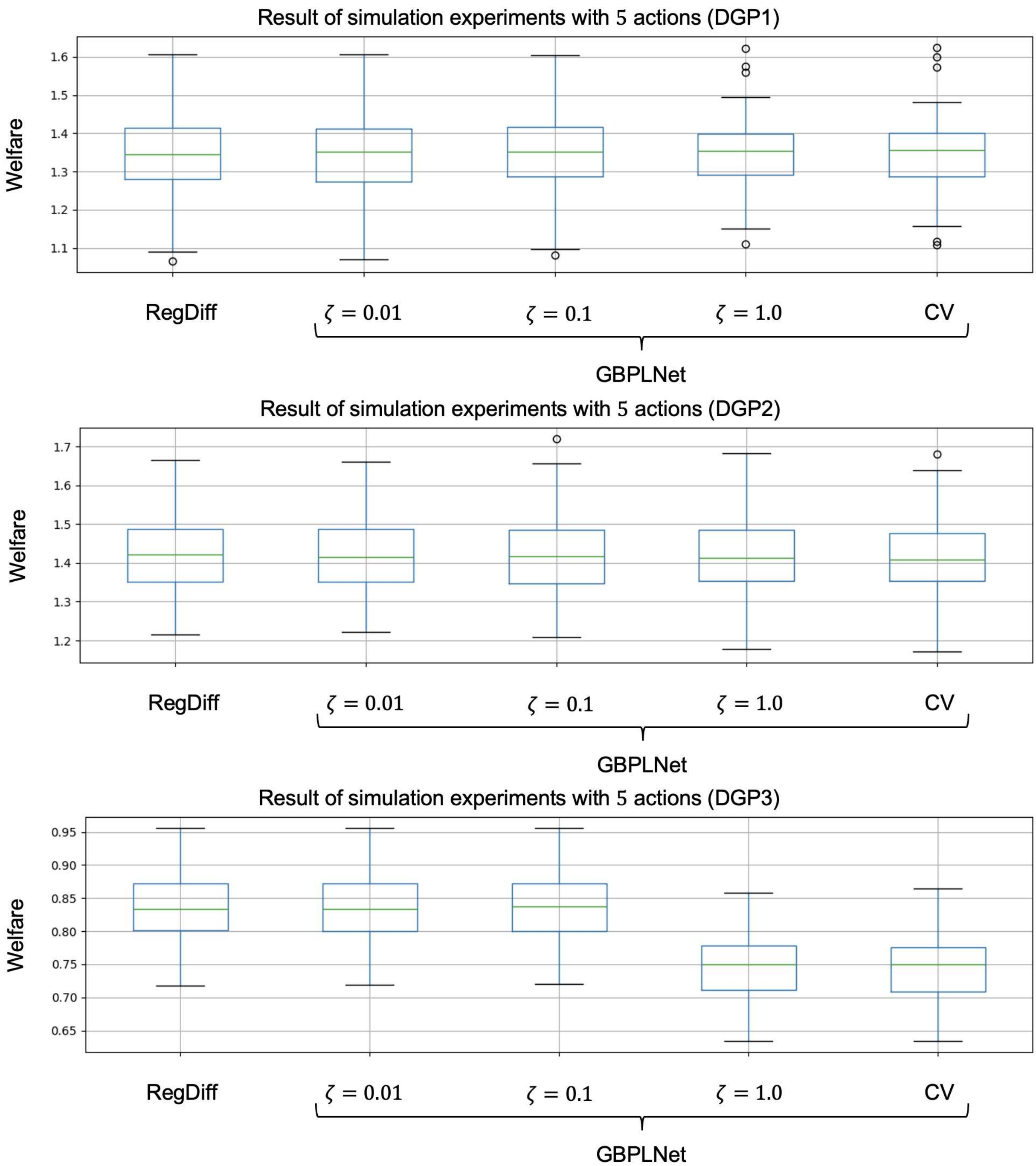

Figure 12: Synthetic $K = 5$ welfare boxplots across 100 trials for DGP1, DGP2, and DGP3.

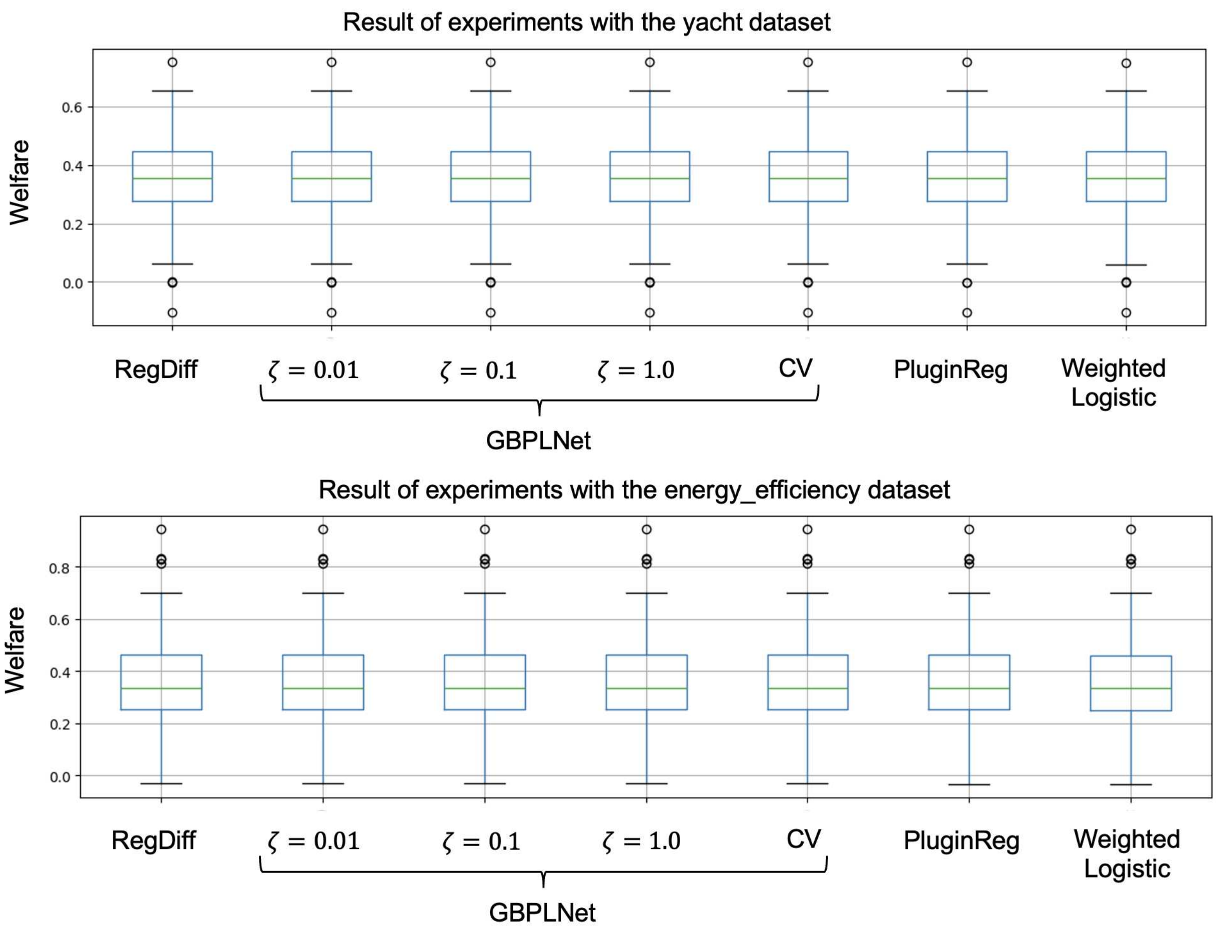

Figure 13: UCI and OpenML welfare boxplots across 100 trials for yacht and energy_efficiency.

Table 12: Counterfactual synthetic $K = 5$ results, 50 trials. Columns report welfare mean, welfare variance, welfare standard error, regret mean, and regret standard error across trials. GBPL-full-DR and GBPL-full-IPW use DR and IPW pseudo-outcome vectors, respectively.

DGP	method	welfare_mean	welfare_var	welfare_se	regret_mean	regret_se
DGP1	GBPL-full-DR ($\zeta = 0.01$)	1.3312	0.0112	0.0150	0.5122	0.0081
DGP1	GBPL-full-DR ($\zeta = 0.1$)	1.3467	0.0127	0.0159	0.4967	0.0084
DGP1	GBPL-full-DR ($\zeta = 1.0$)	1.3313	0.0124	0.0157	0.5122	0.0079
DGP1	GBPL-full-DR (CV)	1.3325	0.0125	0.0158	0.5109	0.0082
DGP1	GBPL-full-IPW ($\zeta = 0.01$)	1.3081	0.0130	0.0161	0.5353	0.0096
DGP1	GBPL-full-IPW ($\zeta = 0.1$)	1.3139	0.0125	0.0158	0.5296	0.0084
DGP1	GBPL-full-IPW ($\zeta = 1.0$)	1.2927	0.0137	0.0165	0.5507	0.0094
DGP1	GBPL-full-IPW (CV)	1.2959	0.0138	0.0166	0.5475	0.0092
DGP1	PluginRegK	1.4011	0.0108	0.0104	0.4424	0.0048
DGP2	GBPL-full-DR ($\zeta = 0.01$)	1.4595	0.0098	0.0140	0.4790	0.0074
DGP2	GBPL-full-DR ($\zeta = 0.1$)	1.4601	0.0092	0.0135	0.4784	0.0072
DGP2	GBPL-full-DR ($\zeta = 1.0$)	1.4347	0.0102	0.0143	0.5037	0.0062
DGP2	GBPL-full-DR (CV)	1.4360	0.0107	0.0146	0.5025	0.0064
DGP2	GBPL-full-IPW ($\zeta = 0.01$)	1.4240	0.0108	0.0147	0.5145	0.0085
DGP2	GBPL-full-IPW ($\zeta = 0.1$)	1.4187	0.0111	0.0149	0.5198	0.0088
DGP2	GBPL-full-IPW ($\zeta = 1.0$)	1.3951	0.0117	0.0153	0.5434	0.0083
DGP2	GBPL-full-IPW (CV)	1.3976	0.0109	0.0148	0.5409	0.0081
DGP2	PluginRegK	1.5138	0.0093	0.0096	0.4247	0.0034
DGP3	GBPL-full-DR ($\zeta = 0.01$)	0.8323	0.0028	0.0075	0.5194	0.0062
DGP3	GBPL-full-DR ($\zeta = 0.1$)	0.8302	0.0029	0.0076	0.5216	0.0064
DGP3	GBPL-full-DR ($\zeta = 1.0$)	0.7224	0.0028	0.0074	0.6293	0.0061
DGP3	GBPL-full-DR (CV)	0.7186	0.0027	0.0074	0.6331	0.0061
DGP3	GBPL-full-IPW ($\zeta = 0.01$)	0.8287	0.0033	0.0081	0.5230	0.0068
DGP3	GBPL-full-IPW ($\zeta = 0.1$)	0.8227	0.0036	0.0085	0.5291	0.0073
DGP3	GBPL-full-IPW ($\zeta = 1.0$)	0.7124	0.0030	0.0077	0.6394	0.0064
DGP3	GBPL-full-IPW (CV)	0.7134	0.0031	0.0078	0.6383	0.0064
DGP3	PluginRegK	0.8177	0.0030	0.0055	0.5341	0.0047

Table 13: Counterfactual UCI and OpenML results, binary actions, 30 trials. Columns report welfare mean, welfare variance, welfare standard error, regret mean, and regret standard error across trials.

dataset	method	welfare_mean	welfare_var	welfare_se	regret_mean	regret_se
energy_efficiency	GBPLNet-DR ($\zeta = 0.01$)	0.3687	0.0422	0.0375	0.0020	0.0004
energy_efficiency	GBPLNet-DR ($\zeta = 0.1$)	0.3688	0.0422	0.0375	0.0019	0.0004
energy_efficiency	GBPLNet-DR ($\zeta = 1.0$)	0.3684	0.0424	0.0376	0.0023	0.0004
energy_efficiency	GBPLNet-DR (CV)	0.3680	0.0423	0.0376	0.0026	0.0005
energy_efficiency	GBPLNet-IPW ($\zeta = 0.01$)	0.3399	0.0462	0.0393	0.0307	0.0066
energy_efficiency	GBPLNet-IPW ($\zeta = 0.1$)	0.3419	0.0479	0.0400	0.0288	0.0062
energy_efficiency	GBPLNet-IPW ($\zeta = 1.0$)	0.3394	0.0471	0.0396	0.0313	0.0050
energy_efficiency	GBPLNet-IPW (CV)	0.3387	0.0494	0.0406	0.0319	0.0065
energy_efficiency	PluginReg	0.3674	0.0418	0.0264	0.0033	0.0003
yacht	GBPLNet-DR ($\zeta = 0.01$)	0.3690	0.0245	0.0286	0.0032	0.0009
yacht	GBPLNet-DR ($\zeta = 0.1$)	0.3684	0.0244	0.0285	0.0038	0.0009
yacht	GBPLNet-DR ($\zeta = 1.0$)	0.3683	0.0243	0.0285	0.0040	0.0007
yacht	GBPLNet-DR (CV)	0.3690	0.0244	0.0285	0.0033	0.0006
yacht	GBPLNet-IPW ($\zeta = 0.01$)	0.3189	0.0257	0.0293	0.0534	0.0080
yacht	GBPLNet-IPW ($\zeta = 0.1$)	0.3095	0.0270	0.0300	0.0628	0.0083
yacht	GBPLNet-IPW ($\zeta = 1.0$)	0.3051	0.0254	0.0291	0.0671	0.0093
yacht	GBPLNet-IPW (CV)	0.3082	0.0261	0.0295	0.0641	0.0096
yacht	PluginReg	0.3659	0.0230	0.0196	0.0064	0.0012