PENN INSTITUTE for ECONOMIC RESEARCH
UNIVERSITY of PENNSYLVANIA

The Ronald O. Perelman Center for Political Science and Economics (PCPSE)
133 South 36th Street
Philadelphia, PA 19104-6297

pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper
20-038

Robust Forecasting

TIMOTHY CHRISTENSEN
New York University

HYUNGSIK ROGER MOON
University of Southern California
Schaeffer Center, and Yonsei University

FRANK SCHORFHEIDE
University of Pennsylvania
CEPR, NBER, PIER

November 23, 2020

https://ssrn.com/abstract=3737629

Robust Forecasting

Timothy Christensen
New York University

Hyungsik Roger Moon
Univ. of Southern California
Schaeffer Center, and Yonsei University

Frank Schorfheide*
University of Pennsylvania
CEPR, NBER, and PIER

This Version: November 23, 2020

Abstract

We use a decision-theoretic framework to study the problem of forecasting discrete outcomes when the forecaster is unable to discriminate among a set of plausible forecast distributions because of partial identification or concerns about model misspecification or structural breaks. We derive "robust" forecasts which minimize maximum risk or regret over the set of forecast distributions. We show that for a large class of models including semiparametric panel data models for dynamic discrete choice, the robust forecasts depend in a natural way on a small number of convex optimization problems which can be simplified using duality methods. Finally, we derive "efficient robust" forecasts to deal with the problem of first having to estimate the set of forecast distributions and develop a suitable asymptotic efficiency theory.

JEL CLASSIFICATION: C11, C14, C23, C53

KEY WORDS: Statistical Decision Theory, Dynamic Discrete Choice, Forecasting, Identification, Minimax Loss, Minimax Regret, Panel Data Models, Robustness, Structural Breaks.

*Correspondence: T. Christensen: Department of Economics, New York University, 19 West 4th Street, 6th floor, New York, NY 10012. Email: timothy.christensen@nyu.edu. H.R. Moon: Department of Economics, University of Southern California, KAP 300, Los Angeles, CA 90089. Email: moonr@usc.edu. F. Schorfheide: Department of Economics, 133 S. 36th Street, University of Pennsylvania, Philadelphia, PA 19104-6297. Email: schorf@ssc.upenn.edu.
We are grateful for comments and suggestions from the participants of the 2019 USC INET Panel Data Forecasting Conference, the 2020 Econometric Society World Congress, the Penn Econometrics Lunch seminar, and the Yale Econometrics seminar. We thank Zhan Gao and Boyuan Zhang for proofreading the manuscript. This material is based upon work supported by the National Science Foundation under Grants No. SES-1919034 (Christensen), SES-1625586 (Moon), and SES-1851634 (Schorfheide).

1 Introduction

In this paper, we study the problem of forecasting discrete outcomes when the researcher is unable to discriminate among a set of plausible forecast distributions. There are several reasons why the forecaster might face uncertainty about the forecast distribution. A leading case is partial identification, in which the data up to the forecast origin only set-identify a subset of parameters of the forecasting model. Uncertainty about the forecast distribution can also arise when the forecaster expands the set of models to accommodate concerns about model misspecification or structural breaks between the in-sample and forecast period.

Suppose that a subset of parameters of a forecasting model are only set-identified. Should the lack of point identification be a concern for the forecaster? At first glance the answer appears to be "no": if the parameters in the identified set generate different forecasts and some of these forecasts are less accurate than others, then we should be able to discriminate among the parameters based on the observed data. To the extent that we are unable to do so, the parameterizations should be observationally equivalent and therefore generate the same forecasts. This intuition is confirmed in the context of vector autoregressions (VARs): while the structural form of the VAR may only be set-identified, forecasts only utilize the reduced form of the VAR which is directly identifiable. This intuition is also confirmed in the context of dynamic linear factor models. The parameters are only identified up to a particular normalization of the latent factors, but each normalization leads to identical forecasts. However, the intuition is wrong in many other important settings.

Our paper makes several contributions. First, we show that the VAR intuition does not apply when forecasting using dynamic discrete choice models for panel data. As is well known (Honoré and Tamer, 2006; Chamberlain, 2010; Chernozhukov, Fernández-Val, Hahn, and Newey, 2013), the homogeneous parameters and the correlated random effects distribution are set-identified when no parametric assumptions are made about the random effects distribution. We demonstrate that in the panel dynamic discrete choice setting different parameters in the identified set lead to different forecasts, some more accurate than others. Unlike a VAR, the panel dynamic discrete choice model has a non-Markovian structure due to the sequential learning about heterogeneous coefficients. As a consequence, parameterizations that are indistinguishable based on a panel of length T may become distinguishable in a panel of length $T+1$.

Second, we construct forecasts that are "robust" to uncertainty about the parameterization θ of the forecasting model among a set of parameterizations Θ_0 that are observationally equivalent at the forecast origin T. We refer to this as uncertainty about the forecast distribution for short. Our robust forecasts minimize either maximum risk or maximum regret (i.e. risk relative to the infeasible Bayes decision under the true forecast distribution) over the set of forecast distributions. We show that for binary (or classification) loss, quadratic loss, and logarithmic loss, the optimal binary

forecast under either robustness criterion depends on two extremum problems which characterize the smallest and largest conditional probabilities for the outcomes being forecast over the set of forecasting distributions. Similarly, robust forecasts in the multinomial case depend in a natural way on a small number of extremum problems. Further, we show that these extremum problems can be simplified by duality arguments for a broad class of models.

Our robust forecasts are not only applicable in settings in which parameters are set-identified, but also in environments in which the forecaster is concerned about model misspecification or there is a structural break at the forecast origin. The common feature is that the forecasts depend on an unknown parameter θ which takes values in a set Θ_0. Under misspecification or structural breaks, the set Θ_0 indexes an enlarged class of models representing plausible deviations from a benchmark model.

Third, we derive "efficient robust" forecasts to deal with the problem of first estimating the set Θ_0 prior to making the forecast. To do so, we express Θ_0 as a (set-valued) function of an identifiable reduced-form parameter P. In order to develop an optimality theory, we evaluate forecasts by their integrated maximum risk or regret, averaging over both P and the data. Under this criterion, the optimal forecast is what we call the Bayesian robust forecast. It is obtained by minimizing the posterior maximum risk or regret which conditions on the data and averages out P based on its posterior distribution.

Fourth, we develop an asymptotic efficiency theory for forecasting discrete outcomes under uncertainty about the parameterization of the forecast distribution when Θ_0 is estimated. We show that in binary and multinomial discrete forecasting problems, forecasts that are asymptotically equivalent to the Bayesian robust forecasts minimize asymptotic integrated maximum risk or regret. We refer to such forecasts as *asymptotically efficient-robust*. We demonstrate that forecasts obtained by replacing P with an efficient first-stage estimator can be strictly dominated by the Bayesian robust forecast. This suboptimality of plug-in forecasts arises in settings in which key statistics that determine the robust forecast are only directionally, but not fully differentiable, with respect to the identifiable reduced-form parameters. Bagged predictors (see Breiman (1996)) that replace posterior averaging with averaging across the bootstrap distribution of an efficient estimator of P, on the other hand, tend to be asymptotically efficient-robust.

Our paper is related to several literatures. For forecasting short time-series using panel data see, e.g., Baltagi (2008), Gu and Koenker (2016), Liu (2019), and Liu, Moon, and Schorfheide (2018, 2020). Applications of partial identification in nonlinear panel data analysis include Honoré and Tamer (2006) and Chernozhukov et al. (2013). Much of our paper is devoted to forecasting binary outcomes which has been previously considered by, for instance, Elliott and Lieli (2013), Lahiri and Yang (2013), and Elliott and Timmermann (2016).

There is an extensive literature on statistical decision theory following Wald (1950). Closely related to our approach are Γ-minimax (or Γ-minimax regret) decisions in robust Bayes analysis (Robbins, 1951; Berger, 1985). In economics, this approach is also related to the multiple priors framework of Gilboa and Schmeidler (1989) and the robustness literature following Hansen and Sargent (2001). For econometric applications, Chamberlain (2000, 2001) derives minimax decision rules under point identification. Kitagawa (2012), Giacomini and Kitagawa (2018), and Giacomini, Kitagawa, and Uhlig (2019) study robust Bayesian analysis under set identification.

Hirano and Wright (2017) study the problem of forecasting continuous outcomes under uncertainty about predictor variables in a weak predictor local asymptotic setting. Despite several differences between their work and ours,[1] they also find that bagging can reduce asymptotic risk. Discrete forecasting has a similar structure to statistical treatment assignment and our efficiency results are related to efficiency results in that literature, most notably Hirano and Porter (2009). In their setting, a welfare contrast is a smooth function of a point-identified, regularly estimable parameter. Their efficient rules are based on plugging-in an efficient estimator of the parameter. In our setting, uncertainty about the forecast distribution can introduce a type of non-smoothness to the robust forecasting problem. In consequence, our efficient robust forecasts differ from plug-in rules.

The remainder of this paper is organized as follows. Section 2 describes the setup, our objectives, and introduces motivating examples. Sections 3 and 4 derive our robust and efficient robust forecasts for binary and multinomial forecasting settings, respectively. Section 3 also contains an application to panel models for dynamic binary choice. Section 5 presents the main results on asymptotic efficiency. Appendix A discusses computation for a broad class of models including semiparametric panel data models. Appendix B contains additional results on robust binary forecasts and all proofs are relegated to Appendix C.

2 Setup, Motivating Examples, and Objectives

2.1 Setup

The econometrician wishes to forecast a random variable Y taking values in a finite set $\mathcal{Y}$. The econometrician assumes Y is distributed according to an (unknown) distribution in a family of forecast distributions $\{\mathbb{P}_\theta(Y = y) : \theta \in \Theta_0\}$, where $\theta \in \Theta$ denotes a vector of parameters and $\Theta_0 \subseteq \Theta$ indexes the set of forecast distributions over which the forecaster seeks robustness. The

[1] For instance, uncertainty about the parameterization of the forecasting model is resolved asymptotically in their framework whereas it persists in our setting.

forecast distributions $\mathbb{P}_\theta(Y = y)$ may be conditioned on covariates observed by the econometrician when making the forecast, but we suppress this dependence in what follows to simplify notation.

As we discussed in the Introduction, there are several reasons why the forecaster might be uncertain about the forecast distribution. A leading case is partial identification, in which Θ_0 represents the identified set of parameters that are observationally equivalent up to the forecast origin. Uncertainty about the forecast distribution can also arise under concerns about model misspecification or structural breaks. In these settings, Θ_0 indexes an enlarged class of models representing plausible deviations from a benchmark model.

2.2 Motivating Examples

To fix ideas and illustrate the broad applicability of our results, we now present a number of examples of where this forecasting problem arises. The first four examples use a panel data model for dynamic discrete choice to show how our approach accommodates concerns about model misspecification and structural breaks in a unified manner, though these are relevant concerns for *any* forecasting model. The last two examples involve counterfactuals and treatment assignments.

Example 1: Semiparametric random effects model for dynamic binary choice. Let

$$Y_{it+1} = \mathbb{I}[\lambda_i + \beta Y_{it} \geq U_{it+1}], \quad \mathbb{P}(U_{it+1} \leq u | Y_i^t = y^t, \lambda_i = \lambda) = \Phi_{t+1}(u), \tag{1}$$

where $\mathbb{I}[y \geq a] = 1$ if $y \geq a$ and 0 otherwise, $Y_i^t = (Y_{i1}, ..., Y_{it})'$, and $y^t \in \{0,1\}^t$. The econometrician observes $Y_i^T = (Y_{i1}, ..., Y_{iT})'$ for $i = 1, \ldots, n$ where T is fixed. To avoid the initial conditions problem, the econometrician treats Y_{i0} as unobserved and specifies a joint distribution $\Pi_{\lambda,y}$ over Y_{i0} and λ_i. As is well known, β is not point-identified when T is small and no parametric restrictions are placed on $\Pi_{\lambda,y}$ (see, e.g., Cox (1958), Chamberlain (1985), and Magnac (2000)). Moreover, $\Pi_{\lambda,y}$ and the Φ_t are not nonparametrically point-identified for any T.

The econometrician wishes to forecast individual-level outcomes Y_{iT+1} conditional upon an individual's history $Y_i^T = y^T \in \{0,1\}^T$. Suppose that the econometrician assumes each of the Φ_t takes a parametric form Φ, such as logistic or standard normal. The identified set Θ_0 then consists of all $(\Pi_{\lambda,y}, \beta)$ for which the model-implied probabilities of observing sequences $Y_i^T = y^T \in \{0,1\}^T$ are equal to the probabilities observed in the data up to date T:

$$\Theta_0 = \left\{\theta = (\beta, \Pi_{\lambda,y}) \in \Theta : p(y^T|\beta, \Pi_{\lambda,y}) = p(y^T) \ \ \forall\, y^T \in \{0,1\}^T\right\}, \tag{2}$$

where $p(y^T|\beta, \Pi_{\lambda,y})$ denotes the model-implied probabilities and $p(y^T)$ denotes the true (population) probabilities of observing $Y_i^T = y^T$. In the above notation, $Y = Y_{iT+1}$ and the forecast probability

$\mathbb{P}_\theta$ denotes the conditional probability over Y_{iT+1} given $Y_i^T = y^T$:

$$\mathbb{P}_\theta(Y = 1) := \mathbb{P}_\theta(Y_{iT+1} = 1 | Y_i^T = y^T) = \frac{\int \Phi(\beta y_{iT} + \lambda) p(y^T | y_0, \lambda; \beta) \mathrm{d}\Pi_{\lambda,y}(\lambda, y_0)}{\int p(y^T | y_0, \lambda; \beta) \mathrm{d}\Pi_{\lambda,y}(\lambda, y_0)} . \quad \square \tag{3}$$

Example 2: Misspecification. Consider the setup described in Example 1, but suppose that the econometrician adopted a parametric correlated random effect model, $\Pi_{\lambda,y} = \Pi(\lambda, y_0; \xi)$ for $\xi \in \Xi$, a set of auxiliary parameters. The econometrician is worried that this parametric random effects specification is misspecified, and so allows for the possibility that $\Pi_{\lambda,y} \in N(\xi)$, a neighborhood of $\Pi(\lambda, y; \xi)$. Suppose the econometrician again sets $\Phi_t = \Phi$ for all t. The set Θ_0 is

$$\Theta_0 = \{\theta = (\beta, \xi, \Pi_{\lambda,y}) \in \Theta : p(y^T | \beta, \Pi_{\lambda,y}) = p(y^T) \ \ \forall\, y^T \in \{0,1\}^T \ \text{ and } \ \Pi_{\lambda,y} \in N(\xi)\} .$$

This setup was considered by Bonhomme and Weidner (2019) under *local* misspecification, where $N(\xi)$ are Kullback–Leibler neighborhoods $N(\xi) = \left\{\Pi : K(\Pi \,\|\, \Pi(\,\cdot\,; \xi)) \leq \delta\right\}$ for each $\xi \in \Xi$ with $\delta \downarrow 0$ as $n \to \infty$, so that worst-case misspecification bias and sampling uncertainty are of the same order asymptotically.[2] We instead treat $\delta > 0$ as fixed, allowing *global* misspecification. $\square$

Example 3: Structural breaks. Three types of breaks can, in principle, occur at the forecast origin T in Example 1: a break in the distribution of the U_{it}, a break in the λ_i, and a break in β. Suppose the econometrician again takes $\Phi_t = \Phi$ for dates $t = 1, \ldots, T$, but allows for the possibility that $\Phi_{T+1} \neq \Phi$. For instance, the econometrician might want to allow for $\Phi_{T+1} \in N$, a neighborhood of Φ. Even if β and $\Pi_{\lambda,y}$ were known at date T, there would still be a set of forecast distributions for Y_{iT+1} corresponding to different $\Phi_{T+1} \in N$. Using the above notation, we can redefine Θ_0 as

$$\Theta_0 = \left\{\theta = (\beta, \Pi_{\lambda,y}, \Phi_{T+1}) \in \Theta : p(y^T | \beta, \Pi_{\lambda,y}) = p(y^T) \ \ \forall\, y^T \in \{0,1\}^T \ \text{ and } \ \Phi_{T+1} \in N\right\},$$

and replace Φ in the definition of $\mathbb{P}_\theta(Y = 1)$ in (3) with Φ_{T+1}. Breaks in λ_i can be viewed as a location shift of the distribution Φ_t and are subsumed under breaks in the distribution of U_{it} for suitable choice of N. Breaks in β can be handled by defining

$$\Theta_0 = \left\{\theta = (\beta, \beta_{T+1}, \Pi_{\lambda,y}) \in \Theta : p(y^T | \beta, \Pi_{\lambda,y}) = p(y^T) \ \ \forall\, y^T \in \{0,1\}^T \ \text{ and } \ |\beta - \beta_{T+1}| \leq \delta\right\},$$

and by replacing $\Phi(\beta y_{iT} + \lambda)$ in (3) with $\Phi(\beta_{T+1} y_{iT} + \lambda)$. $\square$

[2] Note that the emphasis of Bonhomme and Weidner (2019) is on estimating posterior average effects whereas we focus on forecasting discrete (e.g. individual-level) outcomes.

Example 4: Semiparametric random effects model for dynamic multinomial choice. Let

$$Y_{it+1} = \arg\max_m \left(U^m_{it+1} + \varepsilon^m_{it+1}\right)\,, \quad U^0_{it+1} = 0\,, \quad U^m_{it+1} = u_{mt+1}(X_{it}, Y_i^t; \phi, \lambda_i, Y_{i0})\,, \quad m = 1, \ldots, M\,,$$

where $\varepsilon_{it} = (\varepsilon^0_{it}, \ldots, \varepsilon^M_{it})'$ is a vector of utility shocks with $\varepsilon_{it}|X_{it}, \lambda_i \sim \Phi_t$ for each t, Φ_t is a potentially time-varying distribution, ϕ is a vector of homogeneous parameters, λ_i is a vector of heterogeneous parameters, and X_{it} is a vector of exogenous regressors. The econometrician observes $Y_i^T = (Y_{i1}, \ldots, Y_{iT})'$ and $X_i^T = (X_{i1}, \ldots, X_{iT})'$ for $i = 1, \ldots, n$ where T is fixed and $n \to \infty$. To avoid the initial conditions problem, the econometrician specifies a joint distribution $\Pi_{\lambda,y}$ for (λ_i, Y_{i0}). As with Example 1, identification of model parameters with fixed T and $n \to \infty$ is delicate, especially when parametric assumptions about the Φ_t and/or $\Pi_{\lambda,y}$ are relaxed; see Honoré and Kyriazidou (2000), Chernozhukov et al. (2013), Khan, Ouyang, and Tamer (2019), and references therein. Identified sets and forecast probabilities for individual-level outcomes are constructed in a similar manner to Example 1. □

Example 5: Counterfactuals in structural models. Counterfactuals in structural models are also subsumed in our framework when the outcome of interest is discrete, as is often the case for static or dynamic models of discrete choice or discrete games (e.g. firm entry/exit). In the above notation, θ are the structural parameters estimated under one policy regime, the econometrician wishes to predict a variable Y, and the model implies that Y is distributed according to $\mathbb{P}_\theta$ under the intervention. Partial identification can arise on two fronts. First, the model may itself be specified flexibly, leading to a non-singleton identified set Θ_0 of structural parameters. Second, the potential for multiple equilibria and lack of knowledge about an equilibrium selection mechanism under the intervention may lead to a nontrivial set of forecast distributions. This can be subsumed by treating the selection mechanism itself as part of θ, with Θ_0 indexing distributions in a manner that is robust to the type of selection mechanism (see, e.g., Jia (2008), Ciliberto and Tamer (2009), and Grieco (2014)). □

Example 6: Treatment assignment. The problem of making discrete forecasts has a very similar structure to a treatment assignment problem, e.g., determining whether an individual should be vaccinated. Suppose the econometrician has access to a sample of observational data of size n and observes for an individual i the triplet $X_i = \big(D_i, W_{0i}(1 - D_i), W_{1i}D_i\big)$, where D_i is a treatment indicator, and W_0 and W_1 are the potential outcomes ("welfare") of the untreated and the treated individuals. As the sample size tends to infinity, the econometrician is able to estimate the reduced form expectations $P = \big(\mathbb{E}[D], \mathbb{E}[W_0(1 - D)], \mathbb{E}[W_1 D]\big)$.

Let $\mathbb{P}_\theta$ denote the joint distribution of (D, W_0, W_1). To keep the example simple, we assume that the potential outcomes are binary and take values $\{a_0, a_1\}$ and $\{b_0, b_1\}$. Thus, the distribution F_θ is discrete with support on $\{0,1\} \times \{a_0, a_1\} \times \{b_0, b_1\}$. The support points of the potential outcome distribution can be easily point-identified based on two untreated (and two treated) individuals with different outcomes. Thus, we exclude the support points from the definition of θ and P. Using the notation that $\theta_{ijk} = \mathbb{P}\big(D = i, W_0 = a_j, W_1 = b_k\big)$, the identified set $\Theta_0(P)$ is defined by the following set of linear restrictions:

$$\begin{aligned} 0 \leq \theta_{ijk} \leq 1, \qquad & \sum_{i=0,1}\sum_{j=0,1}\sum_{k=0,1} \theta_{ijk} = 1, \quad \mathbb{E}[D] = \sum_{j=0,1}\sum_{k=0,1} \theta_{1jk}, \\ \mathbb{E}[W_0(1-D)] \quad = \quad & \sum_{k=0,1} a_0\theta_{00k} + a_1\theta_{01k}, \quad \mathbb{E}[W_1 D] = \sum_{j=0,1} b_0\theta_{1j0} + b_1\theta_{1j1}, \end{aligned}$$

where θ stacks the θ_{ijk} probabilities.

Define the indicator variable $Y = \mathbb{I}[W_1 \geq W_0]$ which measures whether the treatment effect is (weakly) positive or not. From a policy maker's perspective the treatment effect for an individual not included in the initial trial is uncertain and the treatment decision $d \in \{0,1\}$ can be viewed as a forecast of $Y \in \{0,1\}$ with the understanding that the individual should be treated if the point forecast of Y is one and not treated otherwise.

Dehejia (2005) analyzed this problem in a decision-theoretic framework under point identification with binary treatments. However, the example highlights the well-known result that the distribution of welfare rankings can be partially identified (see, e.g., Manski (1996, 2000) and Heckman, Smith, and Clements (1997)). Manski (2000, 2002, 2004, 2007) used a decision-theoretic framework to analyze optimal treatment in a planning problem under partial identification and advocated minimax and minimax regret approaches.[3] □

2.3 Objectives

We derive two types of forecasts that deal with uncertainty about the forecast distribution. The first are *robust forecasts* which seek to robustify the forecast with respect to Y being distributed according to any distribution in the class $\{\mathbb{P}_\theta : \theta \in \Theta_0\}$. We use minimax and minimax regret criteria as our notion of robustness. The second are *efficient robust forecasts* which deal with the additional problem of having to first estimate Θ_0 from data. The exposition in the remainder of this section and in Section 3 focuses on binary outcomes. We will consider extensions to multinomial outcomes in Section 4.

[3] Manski focuses on population welfare objective whereas here we focus on individual-level outcomes. See also Manski and Tetenov (2007), Hirano and Porter (2009), Tetenov (2012), and Kitagawa and Tetenov (2018) amongst others, for the analysis of treatment rules under social welfare objectives.

Known Θ_0. Given a decision space $\mathcal{D} \subseteq [0,1]$, a loss function $\ell : \{0,1\} \times \mathcal{D} \to \mathbb{R}_+$ and $\theta \in \Theta_0$, the risk of $d \in \mathcal{D}$ under the forecast distribution $\mathbb{P}_\theta$ is

$$\mathbb{E}_\theta[\ell(Y,d)] = \ell(0,d)\,\mathbb{P}_\theta(Y=0) + \ell(1,d)\,\mathbb{P}_\theta(Y=1)\,.$$

The expectation $\mathbb{E}_\theta$ and forecast probabilities $\mathbb{P}_\theta$ may condition on covariates observed by the econometrician when making the forecast; we have suppressed this dependence to simplify notation. The θ-optimal forecast, denoted d^*_θ, minimizes risk under $\mathbb{P}_\theta$:

$$\mathbb{E}_\theta[\ell(Y,d^*_\theta)] = \inf_{d\in\mathcal{D}} \mathbb{E}_\theta[\ell(Y,d)]\,.$$

A *minimax* forecast solves

$$\inf_{d\in\mathcal{D}} \sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\ell(Y,d)]\,. \tag{4}$$

The regret of a forecast is its risk in excess of the risk of the θ-optimal forecast. A *minimax regret* forecast solves

$$\inf_{d\in\mathcal{D}} \sup_{\theta\in\Theta_0} \Big(\mathbb{E}_\theta[\ell(Y,d)] - \mathbb{E}_\theta[\ell(Y,d^*_\theta)]\Big)\,. \tag{5}$$

Robust forecasts are derived under these criteria in Sections 3.2 and 4.2 for binary and multinomial outcomes, respectively.

Estimated Θ_0. In many scenarios the researcher might not know Θ_0 and will therefore need to first estimate the set (or features of Θ_0 germane to the forecasting problem) from a sample of data of size n.[4] Our *efficient robust forecasts* deal with the additional uncertainty that arises from not knowing Θ_0. Here "efficient robust" forecasts are those for which the maximum risk or regret is as close as possible to the maximum risk or regret of an oracle forecast with known Θ_0 (see Section 5). This efficiency notion recognizes that uncertainty about the true identity of the forecast distribution among the class $\{\mathbb{P}_\theta : \theta \in \Theta_0\}$ is the dominant consideration asymptotically, but that estimation error may nevertheless have a material impact on the forecast in any finite sample.

To make the analysis of efficient robust forecasts tractable but reasonably general, we assume that the model for the data, say X_n, and outcome Y is indexed by θ and a k-dimensional vector of reduced-form parameters $P \in \mathcal{P} \subseteq \mathbb{R}^k$. The parameters θ and P are linked by a known mapping $P \mapsto \Theta_0(P)$. For partially identified forecasting models, $\Theta_0(P)$ denotes the identified set if P was the true reduced form parameter value. We assume that X_n and Y are related to θ and P in the

[4] The known-Θ_0 case can be viewed as the limit as $n \to \infty$.

following manner:

$$\mathbb{P}_\theta(Y = y|X_n, P) = \mathbb{P}_\theta(Y = y)\,, \tag{6}$$

$$X_n|\theta, P \sim F_{n,P}\,. \tag{7}$$

Condition (6) implies that Y does not depend on the data X_n or P beyond dependence through θ. Condition (7) says that the distribution of the data is fully summarized by P, which is standard for estimation and inference under partial identification; see, e.g., Moon and Schorfheide (2012). Examples 1-6 can be shown to fit this framework. Here we just discuss Example 1 for brevity.

Example 1 (continued). In this example, the econometrician observes data $X_n = (Y_i^T)_{i=1}^n$. The data are used to estimate the vector $P = (p(y^T))_{y^T \in \{0,1\}^T}$, which collects the probabilities of observing sequences $y^T \in \{0,1\}^T$. The mapping $\Theta_0(P)$ from P to $\theta = (\beta, \Pi_{\lambda,y})$ is defined in (2). Given a history y^T, the distribution of $Y \equiv Y_{iT+1}$ is fully summarized by θ; see (3). Moreover, the distribution of the data is itself multinomial over the different realizations of y^T with probabilities P. Thus $F_{n,P}$ is the product of n multinomial distributions with probabilities P. In this model y_0 and λ are unit specific and the forecasts are constructed based on the posterior distribution of (y_0, λ) conditional on $\theta = (\beta, \Pi_{\lambda,y})$, see (3). □

3 Binary Forecasts

In this section we consider the binary forecasting problem. First, in Section 3.1 we review several common binary loss functions and their corresponding θ-optimal forecasts. In Section 3.2 we derive forecasts that are robust to uncertainty about the forecast distribution and in Section 3.3 we construct efficient robust forecasts for the case of an estimated Θ_0. The forecasts are summarized in Table 1 below. Section 3.4 presents an application to semiparametric panel data models for dynamic binary choice.

3.1 θ-Optimal Forecasts

We consider three loss functions to evaluate forecast accuracy: binary (or classification) loss, quadratic loss, and log predictive probability score.

Binary (or Classification) Loss. The binary loss function for $\mathcal{D} = \{0,1\}$ is

$$\ell_b\big(y, d\big) = a_{10}\mathbb{I}[y = 1,\, d = 0] + a_{01}\mathbb{I}[y = 0,\, d = 1]\,, \tag{8}$$

where $a_{10}, a_{01} \geq 0$. A special case with $a_{10} = a_{01}$ is classification loss $\ell_b(y, d) = \mathbb{I}[y \neq d]$.[5] The θ-optimal forecast is

$$d^*_{b,\theta} = \mathbb{I}\left[\mathbb{P}_\theta(Y = 1) \geq \tfrac{a_{01}}{a_{01}+a_{10}}\right] \tag{9}$$

and its risk is

$$a_{10}\mathbb{P}_\theta(Y = 1) \wedge a_{01}\mathbb{P}_\theta(Y = 0)\,, \tag{10}$$

where $a \wedge b = \min\{a, b\}$. The θ-optimal forecast is not unique when $\mathbb{P}_\theta(Y = 1) = \frac{a_{01}}{a_{01}+a_{10}}$. In this case, however, all θ-optimal forecasts differ only in their handling of ties and have the same risk.

Quadratic Loss. The quadratic loss for $d \in \mathcal{D} = [0, 1]$ is

$$\ell_q\big(y, d\big) = \big(y - d\big)^2\,. \tag{11}$$

With $\mathcal{D} = [0, 1]$ the θ-optimal forecast is the mean of Y under the forecast distribution:

$$d^*_{q,\theta} = \mathbb{E}_\theta[Y] = \mathbb{P}_\theta(Y = 1)\,. \tag{12}$$

Log Loss. Here the loss function for $\mathcal{D} = [0, 1]$ is

$$\ell_p\big(y, d\big) = -\mathbb{I}[y = 1]\log d - \mathbb{I}[y = 0]\log(1 - d)\,. \tag{13}$$

The θ-optimal forecast is also the mean:

$$d^*_{p,\theta} = \mathbb{P}_\theta(Y = 1)\,. \tag{14}$$

Although the θ-optimal forecasts under quadratic loss and log loss are the same, their risks are different: the risk under quadratic loss is the variance of the forecast distribution, whereas the risk under log loss is the entropy of the forecast distribution.

3.2 Robust Forecasts

We now relax the assumption that the forecast distribution $\mathbb{P}_\theta$ is known and derive forecasts that are robust with respect to $\mathbb{P}_\theta$ being any member of the set of forecast distributions $\{\mathbb{P}_\theta : \theta \in \Theta_0\}$. Note, however, that in this section we treat Θ_0 as known.

[5] When $a_{10} = a_{01}$, it is without loss of generality to normalize their common value to 1.

The minimax and minimax regret forecasts will depend on the lower and upper values of the forecast probabilities as θ varies over Θ_0:

$$p_L := \inf_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = 1)\,, \text{ and} \tag{15}$$

$$p_U := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = 1)\,. \tag{16}$$

Although our characterizations are general, the challenge in implementing robust forecasts is to solve these extremum problems which will typically require exploiting some additional structure. Appendix A shows how duality methods may be used to simplify computation in a broad class of models that includes, but is not limited to, semiparametric dynamic binary choice models.

3.2.1 Minimax Forecasts

Binary (or Classification) Loss. We first derive the forecast that solves (4) for the binary loss function ℓ_b from (8) and decision space $\mathcal{D} = \{0, 1\}$. The maximum risk of $d \in \{0, 1\}$ is

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\ell_b(Y, d)] = \begin{cases} a_{01}(1 - p_L) & \text{if } d = 1\,, \\ a_{10}\, p_U & \text{if } d = 0\,. \end{cases} \tag{17}$$

The minimax forecast for binary (or classification) loss is therefore

$$d_{b,mm} = \mathbb{I}\left[a_{01} \leq a_{01} p_L + a_{10} p_U\right] \tag{18}$$

and the minimax risk is

$$\mathcal{R}^*_{b,mm} = (a_{01} - a_{01} p_L) \wedge (a_{10} p_U)\,.$$

The minimax binary forecast is not unique when $a_{01} = a_{01} p_L + a_{10} p_U$. In this case, each minimax forecast differs only in its handling of ties and has the same maximum risk.

Quadratic Loss. We now derive the forecast that solves (4) for the quadratic loss function ℓ_q from (11) and decision space $\mathcal{D} = [0, 1]$. The maximum risk of $d \in [0, 1]$ is

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\ell_q(Y, d)] = \begin{cases} p_U(1 - 2d) + d^2 & \text{if } d < \frac{1}{2}\,, \\ p_L(1 - 2d) + d^2 & \text{if } d > \frac{1}{2}\,, \\ \frac{1}{4} & \text{if } d = \frac{1}{2}\,. \end{cases} \tag{19}$$

The minimax forecast is therefore

$$d_{q,mm} = \begin{cases} p_U & \text{if } p_U \leq \frac{1}{2}\,, \\ p_L & \text{if } p_L \geq \frac{1}{2}\,, \\ \frac{1}{2} & \text{otherwise}\,, \end{cases} \tag{20}$$

and the minimax risk is

$$\mathcal{R}^*_{q,mm} = \begin{cases} p_U(1-p_U) & \text{if } p_U \leq \frac{1}{2}\,, \\ p_L(1-p_L) & \text{if } p_L \geq \frac{1}{2}\,, \\ \frac{1}{4} & \text{otherwise}\,. \end{cases}$$

Log Loss. The minimax forecast $d_{q,mm}$ is also minimax for the log loss function ℓ_p from (13) and decision space $\mathcal{D} = [0,1]$; see Appendix B.

3.2.2 Minimax Regret Forecasts

Binary (or Classification) Loss. We first derive the forecast that solves (5) for the binary loss function from (8) and decision space $\mathcal{D} = \{0,1\}$. In view of (10), the inner maximization problem in (5) becomes

$$\sup_{\theta\in\Theta_0}\Big(\mathbb{E}_\theta[\ell_b(Y,d)] - a_{10}\mathbb{P}_\theta(Y=1) \wedge a_{01}\mathbb{P}_\theta(Y=0)\Big) = \begin{cases} (a_{01} - (a_{01}+a_{10})\,p_L)_+ & \text{if } d=1, \\ ((a_{01}+a_{10})\,p_U - a_{01})_+ & \text{if } d=0, \end{cases} \tag{21}$$

where $a_+ = \max\{a, 0\}$. Therefore, the forecast

$$d_{b,mmr} = \mathbb{I}\left[\left(\tfrac{a_{01}}{a_{01}+a_{10}} - p_L\right)_+ \leq \left(p_U - \tfrac{a_{01}}{a_{01}+a_{10}}\right)_+\right] \tag{22}$$

minimizes maximum regret, and its maximum regret is

$$\mathcal{R}^*_{b,mmr} = (a_{01} - (a_{01}+a_{10})p_L)_+ \wedge ((a_{01}+a_{10})p_U - a_{01})_+\,.$$

The minimax and minimax regret binary forecasts for classification loss (i.e., $a_{01} = a_{10}$) are the same; see Appendix B. As with other forecasts for $\mathcal{D} = \{0,1\}$, the minimax regret forecast is not necessarily unique. Non-uniqueness arises whenever $(\frac{a_{01}}{a_{01}+a_{10}} - p_L)_+ = (p_U - \frac{a_{01}}{a_{01}+a_{10}})_+$. If so, each minimax regret forecast has the same maximum regret and differs only in its handling of ties.

Quadratic Loss. We now derive the forecast that solves (5) for the quadratic loss function ℓ_q from (11) and decision space $\mathcal{D} = [0,1]$. By convexity, the maximum regret of $d \in [0,1]$ is

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\ell_q(Y,d)] = \begin{cases} (p_U - d)^2 & \text{if } d \leq \frac{p_L + p_U}{2}\,, \\ (p_L - d)^2 & \text{if } d \geq \frac{p_L + p_U}{2}\,. \end{cases} \tag{23}$$

The minimax regret forecast for quadratic loss is therefore the midpoint of the extreme forecast probabilities:

$$d_{q,mmr} = \frac{p_L + p_U}{2}$$

and the minimax regret is

$$\mathcal{R}^*_{q,mmr} = \left(\frac{p_U - p_L}{2}\right)^2 .$$

Log Loss. Finally, we derive the forecast that solves (5) for the log loss function from (13) and decision space $\mathcal{D} = [0,1]$. The minimax forecast is the θ-optimal forecast if $p_L = p_U$. Suppose $p_L < p_U$. The regret of any $d \in [0,1]$ is the Kullback–Leibler (KL) divergence

$$\mathbb{P}_\theta(Y=1) \log\left(\frac{\mathbb{P}_\theta(Y=1)}{d}\right) + \mathbb{P}_\theta(Y=0) \log\left(\frac{\mathbb{P}_\theta(Y=0)}{1-d}\right) .$$

By convexity, the maximum regret must be obtained at either p_L or p_U:

$$\sup_{\theta \in \Theta_0} \Big(\mathbb{E}_\theta[\ell_p(Y,d)] - \mathbb{E}_\theta[\ell_p(Y,d^*_{p,\theta})]\Big) = \max_{p \in \{p_L, p_U\}} \left(p \log\left(\frac{p}{d}\right) + (1-p)\log\left(\frac{1-p}{1-d}\right)\right) . \tag{24}$$

When $p = p_L$, term in parentheses is increasing for $d \geq p_L$ and when $p = p_U$ the term in parentheses is decreasing for $d \leq p_L$. The maximum regret is therefore minimized by choosing d to equate the two values. The minimax regret forecast under the log-scoring rule $d_{p,mmr}$ uniquely solves

$$\log\left(\frac{d_{p,mmr}}{1 - d_{p,mmr}}\right) = \frac{h(p_U) - h(p_L)}{p_U - p_L}\,, \tag{25}$$

where $h(p) = -p \log p - (1-p)\log(1-p)$ is the entropy of a Bernoulli distribution with success probability p. The minimax regret forecast is therefore the value p that minimizes the maximum KL divergence between the Bernoulli distribution and the forecast distribution $\mathbb{P}_\theta$ over $\theta \in \Theta_0$.

3.3 Efficient Robust Forecasts

We now dispense with the assumption that Θ_0 is known. We consider the setting described at the end of Section 2 in which the econometrician wishes to forecast Y having observed data X_n. In order to develop an optimality theory, we evaluate forecasts by their *integrated maximum risk*,

defined as

$$\mathcal{B}^n_{mm}(d_n;\pi) = \int\int \sup_{\theta\in\Theta_0(P)} \mathbb{E}_\theta[\ell(Y,d(X_n))]\,\mathrm{d}\Pi_n(P|X_n)\,\mathrm{d}F_n(X_n) \tag{26}$$

We will use a similar definition of *integrated maximum regret*, denoted by $\mathcal{B}_{mmr}(d_n;\pi)$. Here, π is a prior distribution for P with support $\mathcal{P}$, $\Pi_n(P|X_n)$ is the posterior distribution of P after having observed the data X_n, and $F_n(X_n)$ is the marginal distribution of the data X_n. As in Section 3.2, conditional on $P\in\mathcal{P}$, we consider the maximum risk or regret over $\Theta_0(P)$. This is the maximum risk faced by the forecaster if P were the true reduced-form parameter. We then average across the joint distribution of (P,X_n) to obtain the integrated maximum risk. The factorization of the joint distribution in the conditional distribution $\Pi_n(P|X_n)$ and the marginal distribution $F_n(X_n)$ highlights the well-known result that the integrated (maximum) risk is minimized by choosing the forecast that minimizes the *posterior (maximum) risk* for each realization X_n. We denote the optimal forecasts under the risk and regret objectives by $d_{b,mm}$ and $d_{b,mmr}$, respectively. We refer to them as *Bayesian robust forecasts* and derive explicit formulas in the remainder of this subsection.

Remark 3.1. While it may seem asymmetric to use an integrated (or Bayes) criterion to deal with P but minimax risk or regret to deal with θ conditional on P, there are are two reasons for doing so. The first is from a robust Bayes perspective on the forecasting problem under partial identification. If the true value P_0 is identified and consistently estimable, then the posterior for P will not depend on π asymptotically. In contrast, the data do not update prior beliefs about θ over the identified set Θ_0. Therefore, the posterior for θ in a Bayesian implementation will depend on the prior asymptotically (see, e.g., Moon and Schorfheide (2012)). Our use of minimax criteria to deal with partial identification of θ can be motivated from robustness considerations with respect to the prior on θ (Kitagawa, 2012; Giacomini and Kitagawa, 2018). The second is practical: average criteria lead to tractable, easily implementable forecasts. □

3.3.1 Minimax Forecasts

We now make dependence of p_L and p_U on the reduced-form parameter explicit by writing

$$p_L(P) := \inf_{\theta\in\Theta_0(P)} \mathbb{P}_\theta(Y=1)\,, \qquad\qquad p_U(P) := \sup_{\theta\in\Theta_0(P)} \mathbb{P}_\theta(Y=1)\,.$$

Binary (or Classification) Loss. In view of (27), the posterior average maximum risk of choosing $d\in\{0,1\}$ is

$$\int \sup_{\theta\in\Theta_0(P)} \mathbb{E}_\theta[\ell_b(Y,d)]\,\mathrm{d}\Pi_n(P|X_n) = \begin{cases} a_{01}\left(1-\int p_L(P)\,\mathrm{d}\Pi_n(P|X_n)\right) & \text{if } d=1\,, \\ a_{10}\int p_U(P)\,\mathrm{d}\Pi_n(P|X_n) & \text{if } d=0\,. \end{cases} \tag{27}$$

The Bayesian robust forecast is therefore

$$d_{b,mm}(X_n) = \mathbb{I}\left[a_{01} \leq \int \left(a_{01} p_L(P) + a_{10} p_U(P)\right) \mathrm{d}\Pi_n(P|X_n)\right]. \tag{28}$$

Quadratic Loss. In view of (19), the posterior average maximum risk of choosing $d \in [0,1]$ is

$$\int \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[\ell_q(Y,d)] \, \mathrm{d}\Pi_n(P|X_n) = \begin{cases} (\int p_U(P) \, \mathrm{d}\Pi_n(P|X_n))(1-2d) + d^2 & \text{if } d < \frac{1}{2}, \\ (\int p_L(P) \, \mathrm{d}\Pi_n(P|X_n))(1-2d) + d^2 & \text{if } d > \frac{1}{2}, \\ \frac{1}{4} & \text{if } d = \frac{1}{2}. \end{cases}$$

The Bayesian robust forecast is therefore

$$d_{q,mm}(X_n) = \begin{cases} \int p_U(P) \, \mathrm{d}\Pi_n(P|X_n) & \text{if } \int p_U(P) \, \mathrm{d}\Pi_n(P|X_n) \leq \frac{1}{2}, \\ \int p_L(P) \, \mathrm{d}\Pi_n(P|X_n) & \text{if } \int p_L(P) \, \mathrm{d}\Pi_n(P|X_n) \geq \frac{1}{2}, \\ \frac{1}{2} & \text{otherwise} \end{cases} \tag{29}$$

Log Loss. The forecast $d_{q,mm}(X_n)$ is also the Bayesian robust forecast for the log loss function ℓ_p from (13) and decision space $\mathcal{D} = [0,1]$; see Appendix B.

3.3.2 Minimax Regret Forecasts

Binary Loss. In view of (21), the posterior average maximum regret for $d \in \{0,1\}$ is

$$\begin{aligned} &\int \sup_{\theta \in \Theta_0(P)} \left(\mathbb{E}_\theta[\ell_b(Y,d)] - a_{10}\mathbb{P}_\theta(Y=1) \wedge a_{01}\mathbb{P}_\theta(Y=0)\right) \mathrm{d}\Pi_n(P|X_n) \\ &= \begin{cases} \left(a_{01} - (a_{01}+a_{10})(\int p_L(P) \, \mathrm{d}\Pi_n(P|X_n))\right)_+ & \text{if } d = 1, \\ \left((a_{01}+a_{10}) \left(\int p_U(P) \, \mathrm{d}\Pi_n(P|X_n)\right) - a_{01}\right)_+ & \text{if } d = 0. \end{cases} \end{aligned}$$

The Bayesian robust forecast is therefore

$$d_{b,mmr}(X_n) = \mathbb{I}\left[\int \left(\tfrac{a_{01}}{a_{01}+a_{10}} - p_L(P)\right)_+ \mathrm{d}\Pi_n(P|X_n) \leq \int \left(p_U(P) - \tfrac{a_{01}}{a_{01}+a_{10}}\right)_+ \mathrm{d}\Pi_n(P|X_n)\right]. \tag{30}$$

Quadratic Loss. In view of (23), the posterior average maximum risk of choosing $d \in [0,1]$ is

$$\int (p_U(P) - d)^2 \mathbb{I}\left[d < \frac{p_L(P) + p_U(P)}{2}\right] + (p_L(P) - d)^2 \mathbb{I}\left[d \geq \frac{p_L(P) + p_U(P)}{2}\right] \mathrm{d}\Pi_n(P|X_n).$$

The Bayesian robust forecast $d_{q,mmr}(X_n)$ is the minimizing value for $d \in [0,1]$, which can be computed numerically (e.g. by replacing the integral with the average across a large number of draws from the posterior then minimizing with respect to d).

Log Loss. In view of (24), the posterior average maximum risk of choosing $d \in [0,1]$ is

$$\int \max_{p \in \{p_L(P), p_U(P)\}} \left(p \log\left(\frac{p}{d}\right) + (1-p) \log\left(\frac{1-p}{1-d}\right) \right) \mathrm{d}\Pi_n(P|X_n)$$

The Bayesian robust forecast $d_{p,mmr}(X_n)$ is the minimizing value for $d \in [0,1]$, which again can be computed numerically.

3.3.3 Summary

Table 1 summarizes the θ-optimal, the robust, and the Bayesian robust decisions under binary, quadratic, and log loss functions.

3.4 Numerical Illustration

We close this section with a numerical illustration to show how uncertainty about the true $\theta \in \Theta_0$ can induce substantial variation in the implied forecast distribution in nonlinear models. We use a panel probit design from Honoré and Tamer (2006). The model is as in Example 1 with Φ_t taken to be the standard normal cdf for all t. The distribution $\Pi_{\lambda,y}$ is unspecified, but λ is assumed to be supported on the discrete evenly-spaced grid $\{-3, -2.8, \ldots, 2.8, 3\}$. Under the true data-generating process, λ and Y_{i0} are independent with Y_{i0} taking the value 0 or 1 with probability $\frac{1}{2}$ and the probability mass for λ is assigned by interpolating a $N(0,1)$ distribution on the support points.[6]

In this example, we wish to forecast Y_{iT+1} having observed Y_i^T. In our earlier notation, the outcome of interest Y represents Y_{iT+1} and the forecast probability $\mathbb{P}_\theta(Y=1)$ denotes the probability under θ that $Y_{iT+1} = 1$ given Y_i^T; see display (3). The identified set Θ_0 consists of all $(\beta, \Pi_{\lambda,y})$ that can match the model-implied probabilities of observing sequences $Y_i^T = y^T$ with the true probabilities for all $y^T \in \{0,1\}^T$; see display (2).

We may compute the set of forecast probabilities $\{\mathbb{P}_\theta(Y=1) : \theta \in \Theta_0\}$ by adapting linear programming methods from Honoré and Tamer (2006) as follows. Denote the support of $\Pi_{\lambda,y}$ by $(\lambda_1, y_{01}), \ldots, (\lambda_L, y_{0L})$. As the support of $\Pi_{\lambda,y}$ is discrete, we identify $\Pi_{\lambda,y}$ with a L-vector $\pi \in \Delta^{L-1} := \left\{x \in \mathbb{R}_+^L : \sum_{l=1}^L x_l = 1\right\}$ and write the restrictions defining Θ_0 in display (2) as $G(\beta)\pi = r$. The matrix $G(\beta)$ is a $2^T \times L$ matrix whose l^{th} column $G_l(\beta)$ is the 2^T-vector of model-implied probabilities of observing different realizations of Y_i^T conditional on $\lambda_i = \lambda_l$ and $Y_{i0} = y_{0l}$:

$$G_l(\beta) = \left(p(y^T | y_{0l}, \lambda_l; \beta) \right)_{y^T \in \{0,1\}^T},$$

[6] See p.619 in Honoré and Tamer (2006) for details.

	θ-optimal:
	$d^*_\theta := \operatorname{argmin}_{d\in\mathcal{D}} \mathbb{E}_\theta[\ell(Y,d)]$
Binary ℓ_b	$d^*_{b,\theta} = \mathbb{I}[(a_{01}+a_{10})\mathbb{P}_\theta(Y=1) \geq a_{01}]$
Quadratic ℓ_q	$d^*_{q,\theta} = \mathbb{P}_\theta(Y=1)$
Log ℓ_p	$d^*_{p,\theta} = d^*_{q,\theta}$
	Robust (minimax):
	$d_{mm} := \operatorname{argmin}_{d\in\mathcal{D}} \left(\sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\ell(Y,d)]\right)$
Binary ℓ_b	$d_{b,mm} = \mathbb{I}\left[a_{01}p_L + a_{10}p_U \geq a_{01}\right]$
Quadratic ℓ_q	$d_{q,mm} = \begin{cases} p_U & \text{if } p_U \leq \frac{1}{2}\,, \\ p_L & \text{if } p_L \geq \frac{1}{2}\,, \\ \frac{1}{2} & \text{otherwise}\,, \end{cases}$
Log ℓ_p	$d_{p,mm} = d_{q,mm}$
	Robust (minimax regret):
	$d_{mmr} := \operatorname{argmin}_{d\in\mathcal{D}} \left(\sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\ell(Y,d)] - \mathbb{E}_\theta[\ell(Y,d^*_\theta)]\right)$
Binary ℓ_b	$d_{b,mmr} = \mathbb{I}[(\frac{a_{01}}{a_{01}+a_{10}} - p_L)_+ \leq (p_U - \frac{a_{01}}{a_{01}+a_{10}})_+]$
Quadratic ℓ_q	$d_{q,mmr} = \frac{1}{2}(p_L + p_U)$
Log ℓ_p	$d_{q,mmr} =$ see equation (25)
	Bayesian robust (minimax):
	$d_{mm}(X_n) := \operatorname{argmin}_{d\in\mathcal{D}} \int \sup_{\theta\in\Theta_0(P)} \mathbb{E}_\theta[\ell(Y,d)]\,\mathrm{d}\Pi_n$
Binary ℓ_b	$d_{b,mm}(X_n) = \mathbb{I}[a_{01} \leq \int (a_{01}p_L(P) + a_{10}p_U(P))\ \mathrm{d}\Pi_n]$
Quadratic ℓ_q	$d_{q,mm}(X_n) = \begin{cases} \int p_U(P)\,\mathrm{d}\Pi_n & \text{if } \int p_U(P)\,\mathrm{d}\Pi_n \leq \frac{1}{2}\,, \\ \int p_L(P)\,\mathrm{d}\Pi_n & \text{if } \int p_L(P)\,\mathrm{d}\Pi_n \geq \frac{1}{2}\,, \\ \frac{1}{2} & \text{otherwise} \end{cases}$
Log ℓ_p	$d_{p,mm}(X_n) = d_{q,mm}(X_n)$
	Bayesian robust (minimax regret)
	$d_{mmr}(X_n) := \operatorname{argmin}_{d\in\mathcal{D}} \int (\sup_{\theta\in\Theta_0(P)} \mathbb{E}_\theta[\ell(Y,d)] - \mathbb{E}_\theta[\ell(Y,d^*_\theta)])\,\mathrm{d}\Pi_n$
Binary ℓ_b	$d_{b,mmr}(X_n) = \mathbb{I}[\int (\frac{a_{01}}{a_{01}+a_{10}} - p_L(P))_+ \mathrm{d}\Pi_n \leq \int (p_U(P) - \frac{a_{01}}{a_{01}+a_{10}})_+ \mathrm{d}\Pi_n]$
Quadratic ℓ_q	$d_{q,mmr}(X_n) = \arg\min_{d\in[0,1]} \int (p_U(P)-d)^2 \mathbb{I}[d < \frac{p_L(P)+p_U(P)}{2}]\,\mathrm{d}\Pi_n$ $+ \int (p_L(P)-d)^2 \mathbb{I}[d \geq \frac{p_L(P)+p_U(P)}{2}]\,\mathrm{d}\Pi_n$
Log ℓ_p	$d_{p,mmr}(X_n) = \arg\min_{d\in[0,1]} \int \max_{p\in\{p_L(P),p_U(P)\}} (p\log(\frac{p}{d}) + (1-p)\log(\frac{1-p}{1-d}))\,\mathrm{d}\Pi_n$

Table 1: Summary of Binary Forecasts. Binary forecasts are for the binary loss function ℓ_b from (8), quadratic loss function ℓ_q from (11), and log loss ℓ_p from (13). To simplify notation, Π_n denotes $\Pi_n(P|X_n)$.

where

$$p(y^T|y_0,\lambda;\beta) = \prod_{t=1}^{T} \Phi(\beta y_{t-1}+\lambda)^{y_t}(1-\Phi(\beta y_{t-1}+\lambda))^{1-y_t}\,, \tag{31}$$

and $r = \left(p(y^T)\right)_{y^T\in\{0,1\}^T}$ is the 2^T-vector that collects the true probabilities of each realization of the sequences Y_i^T. The forecast probability $\mathbb{P}_\theta(Y_{iT+1}=1|Y_i^T=y^T)$ from display (3) may also be

written as $b(\beta)'\pi$ where $b(\beta)$ is an L-vector whose l^{th} entry is

$$b_l(\beta) = \frac{\Phi(\beta y_T + \lambda_l)p(y^T|y_{0l}, \lambda_l; \beta)}{p(y^T)},$$

where y_T denotes the element of y^T corresponding to date T. Note that we can place $p(y^T)$ in the denominator because $\int p(y^T|y_0, \lambda; \beta)\mathrm{d}\Pi_{\lambda,y}(\lambda, y_0) = p(y^T)$ for any $\theta \in \Theta_0$.

The problem of computing p_U can be written as

$$p_U = \sup_{\beta} \left(\sup_{\pi \in \Delta^{L-1}} b(\beta)'\pi \quad \text{s.t.} \quad G(\beta)\pi = r \right),$$

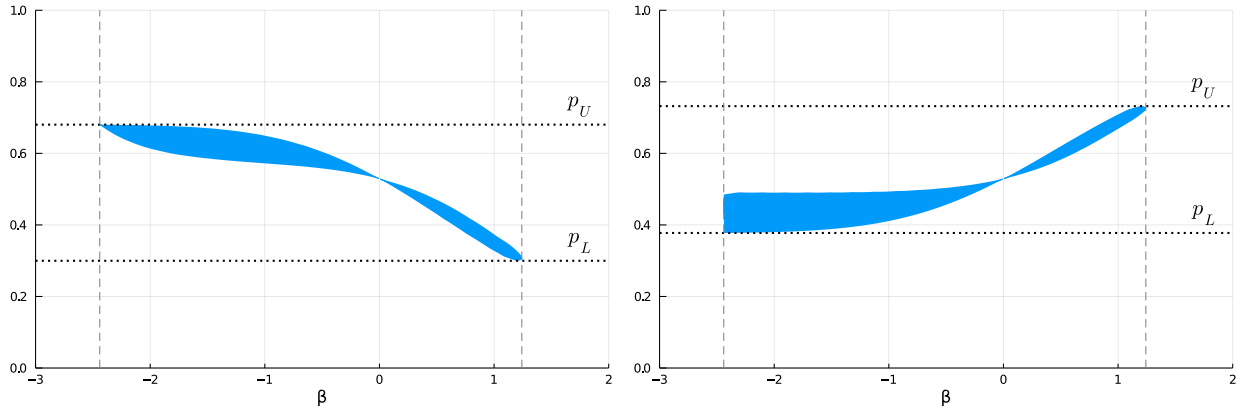
with the understanding that the value of the inner maximization over π is $-\infty$ if there does not exist a π for which $G(\beta)\pi = r$. For such value of β there does not exist a $\Pi_{\lambda,y}$ such that the model can explain the observed probabilities up to date T. As shown in Appendix A, the inner optimization over $\Pi_{\lambda,y}$ can be rewritten as a linear program, leading to the equivalent representation

$$p_U = \sup_{\beta} \left(\inf_{v \in \mathbb{R}^{K+1}} [0_{1\times K}\,,\, 1]\, v \quad \text{s.t.} \quad A(\beta)v \leq -b(\beta) \right), \tag{32}$$

where $K = 2^T$ and $A(\beta) = [G(\beta)' - (1_{L\times 1} \otimes r')\,,\, -1_{L\times 1}]$ with $\otimes$ denoting the Kronecker product. The lower value is computed similarly; see Appendix A.

Suppose $T = 2$ and the true $\beta_0 = 0.2$. The identified set for β is approximately $[-2.4403, 1.2428]$; these are all values of β for which there is a $\Pi_{\lambda,y}$ such that the model can explain the observed probabilities up to date $T = 2$. The limits of the identified set for β are denoted as grey dashed vertical lines in Figure 1. For each value of β in this set, we compute the smallest and largest values of the forecast probability $\mathbb{P}_\theta(Y = 1)$ subject to the constraint that $(\beta, \Pi_{\lambda,y}) \in \Theta_0$. The linear programming problem to compute the upper probability conditional on β is characterized in parentheses on the right-hand side of (32). The range of forecast probabilities as a function of β is shown in Figure 1. Maximizing and minimizing with respect to β yields the values p_L and p_U; these are marked as black dotted horizontal lines in Figure 1. Although each $(\beta, \Pi_{\lambda,y}) \in \Theta_0$ induces identical distributions over Y_i^T, they induce very different distributions over Y_{iT+1} and hence generate very different θ-optimal forecasts. As a consequence, some parameterizations that are indistinguishable based on T observations, become distinguishable based on $T+1$ observations and the identified set shrinks over time. This feature is due to the sequential learning about heterogeneous parameters that generates a non-Markovian structure of the model.

In this numerical example, we consider robust forecasts which take Θ_0 as known. This is the asymptotic problem faced by the forecaster in a large-n, fixed-T setting. Suppose we condition on



(a) $Y_{iT} = 0$ (b) $Y_{iT} = 1$

Figure 1: Panel probit example with $T = 2$ and $\beta_0 = 0.2$. Shaded regions denote the sets $\{\mathbb{P}_\theta(Y = 1) : (\beta, \Pi_{\lambda,y}) \in \Theta_0\}$ as a function of β. Black dotted lines denote p_L and p_U.

Y_i^T with $Y_{iT} = 0$.[7] The set of forecast probabilities $\{\mathbb{P}_\theta(Y = 1) : \theta \in \Theta_0\}$ is wide, spanning from $p_L = 0.2997$ to $p_U = 0.6803$ (see the left panel of Figure 1). In particular, there are $\theta \in \Theta_0$ for which $\mathbb{P}_\theta(Y = 1) < \frac{1}{2}$ so the θ-optimal decision would be $d^*_{b,\theta} = 0$ for these θ. However, there are other $\theta \in \Theta_0$ for which $\mathbb{P}_\theta(Y = 1) > \frac{1}{2}$ and therefore the corresponding θ-optimal decision would be $d^*_{b,\theta} = 1$ for these θ. Our robust forecasts are useful here as the forecaster has no way to discriminate among $\theta \in \Theta_0$ based on date-T information. As $p_L + p_U < 1$, the minimax and minimax regret forecast for symmetric binary loss is therefore $d_{b,mm} = d_{b,mmr} = 0$. Similarly, when $Y_{iT} = 1$ the set of forecast probabilities $\{\mathbb{P}_\theta(Y = 1) : \theta \in \Theta_0\}$ is again quite wide, spanning $p_L = 0.3775$ to $p_U = 0.7320$ (see the right panel of Figure 1). Here $p_L + p_U > 1$ so $d_{b,mm} = d_{b,mmr} = 1$.

4 Multinomial Forecasts

We now extend the preceding analysis to multinomial forecasts. We first describe θ-optimal forecasts with known θ (Subsection 4.1), then describe forecasts that are robust with respect to the set of forecasting models $\{\mathbb{P}_\theta : \theta \in \Theta_0\}$ with Θ_0 known (Subsection 4.2), before concluding with efficient robust forecasts that deal with both model and sampling uncertainty (Subsection 4.3). The forecasts are summarized in Table 2 at the end of this section.

4.1 θ-optimal Forecasts

Throughout this section we focus on *classification loss* for the decision space $\mathcal{D} = \{0, 1, \ldots, M\}$:

$$\ell_c(y, d) = \mathbb{I}[y \neq d]\,. \tag{33}$$

[7] In this design, the conditional distribution of Y_{iT+1} given Y_i^T depends only on Y_{iT}.

This loss function generalizes binary loss in the symmetric case (i.e., $a_{01} = a_{10}$) to multinomial forecasts.[8] The θ-optimal forecast in this environment under $\mathbb{P}_\theta$ is the most likely outcome:

$$d^*_{c,\theta} \in \arg\max_m \mathbb{P}_\theta(Y = m)\,. \tag{34}$$

In the above display we write "$\in$" to allow for the possibility of ties. When the arg max is non-singleton, any element of the set of minimizers is a θ-optimal point forecast. Any θ-optimal forecast has risk

$$1 - \max_m \mathbb{P}_\theta(Y = m)\,.$$

4.2 Robust Forecasts

We now derive the minimax and minimax regret forecasts that solve the decision problems (4) and (5) for the classification loss function ℓ_c from (33) and decision space $\mathcal{D} = \{0, 1, \ldots, M\}$.

4.2.1 Minimax Forecasts

In the multivariate case, the analogues of p_L and p_U are the $M + 1$ quantities

$$\underline{p}_m = \inf_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = m)\,, \quad m \in \{0, 1, \ldots, M\}\,. \tag{35}$$

Computation of $\underline{p}_m$ using duality methods is discussed in Appendix A. The maximum risk from choosing $d \in \mathcal{D}$ is

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\ell_c(Y, d)] = 1 - \underline{p}_d\,. \tag{36}$$

The minimax forecast for classification loss is therefore

$$d_{c,mm} \in \arg\max_m \underline{p}_m \tag{37}$$

and the minimax risk is

$$\mathcal{R}^*_{c,mm} = 1 - \max_m \underline{p}_m\,. \tag{38}$$

As before, the minimax-optimal forecast is not necessarily unique. Non-uniqueness arises when the set of maximizers of $m \mapsto \underline{p}_m$ is not a singleton. If so, each minimax-optimal forecast differs only in its handling of ties and has the same maximum risk.

[8] It is straightforward to modify what follows to penalize some types misclassifications more heavily than others, as we did in the binary case. We adopt the equal-weighted specification (33) for notational convenience.

4.2.2 Minimax Regret Forecasts

For minimax regret forecasts, define

$$\Delta p_m := \sup_{\theta \in \Theta_0} \left(\max_{m'} \mathbb{P}_\theta(Y = m') - \mathbb{P}_\theta(Y = m) \right) . \tag{39}$$

Suppose the forecaster chooses d. The difference $\max_{m'} \mathbb{P}_\theta(Y = m') - \mathbb{P}_\theta(Y = d)$ is the regret from this choice under the forecast distribution $\mathbb{P}_\theta$. Having chosen d, the quantity Δp_d is therefore the forecaster's maximum regret over all $\theta \in \Theta_0$. The minimax regret forecast is therefore

$$d_{c,mmr} \in \arg\min_m \Delta p_m .$$

and the minimax regret is

$$\mathcal{R}^*_{c,mmr} = \min_m \Delta p_m .$$

Unlike the binary case, equivalence of minimax and minimax regret forecasts for classification loss no longer holds when $M \geq 2$; see Appendix B. Computation of Δp_m is discussed in Appendix A.

4.3 Efficient Robust Forecasts

In this section we now drop the assumption that Θ_0 is known and consider also the need to estimate features of Θ_0 that are relevant for the forecasting problem from data. We consider the same setup and notation as developed for the binary case in Section 3.3.

4.3.1 Minimax Forecasts

Here we make dependence on the reduced-form parameter explicit by defining

$$\underline{p}_m(P) := \inf_{\theta \in \Theta_0(P)} \mathbb{P}_\theta(Y = m) .$$

In view of (36), the posterior average maximum risk of choosing $d \in \mathcal{D}$ is

$$\int \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[\ell_c(Y, d)] \, \mathrm{d}\Pi_n(P|X_n) = 1 - \int \underline{p}_d(P) \, \mathrm{d}\Pi_n(P|X_n)$$

The Bayesian robust forecast is therefore

$$d_{c,mm}(X_n) \in \arg\max_m \left(\int \underline{p}_m(P) \, \mathrm{d}\Pi_n(P|X_n) \right) . \tag{40}$$

θ-optimal	$d^*_{c,\theta} \in \arg\max_m \mathbb{P}_\theta(Y = m)$
Robust (minimax)	$d_{c,mm} \in \arg\min_m \Delta p_m$
Robust (minimax regret)	$d_{c,mmr} \in \arg\min_m \Delta p_m$
Bayesian robust (minimax)	$d_{c,mm}(X_n) \in \arg\max_m \left(\int \underline{p}_m(P) \, \mathrm{d}\Pi_n(P\|X_n) \right)$
Bayesian robust (minimax regret)	$d_{c,mmr}(X_n) \in \arg\min_m \left(\int \Delta p_m(P) \, \mathrm{d}\Pi_n(P\|X_n) \right)$

Table 2: Summary of Multinomial Forecasts. Multinomial forecasts are for the classification loss function ℓ_c from (33).

In case of ties, any (possibly randomized) tie-breaking rule is optimal. For instance, one could simply choose the smallest value of m among the set of maximizers.

4.3.2 Minimax Regret Forecasts

For minimax regret forecasts, define

$$\Delta p_m(P) := \sup_{\theta \in \Theta_0(P)} \left(\max_{m'} \mathbb{P}_\theta(Y = m') - \mathbb{P}_\theta(Y = m) \right).$$

The posterior average maximum regret of choosing $d \in \mathcal{D}$ is

$$\int \Delta_{p_d}(P) \, \mathrm{d}\Pi_n(P|X_n).$$

The Bayesian robust forecast is therefore

$$d_{c,mmr}(X_n) \in \arg\min_m \left(\int \Delta p_m(P) \, \mathrm{d}\Pi_n(P|X_n) \right). \tag{41}$$

In case of ties, any (possibly randomized) tie-breaking rule is optimal. For instance, one could simply choose the smallest value of m among the set of minimizers.

5 Asymptotic Efficiency for the Robust Forecasting Problem

In this section we focus on forecasts that are asymptotically efficient-robust. We continue to evaluate the forecasts by their *integrated maximum risk* (or regret), but only require this criterion to be minimized in the limit as the sample size n tends to infinity. This enlarges the class of efficient forecasts to those that are asymptotically equivalent to the Bayesian robust forecast.

Some interesting findings emerge. First, "plug-in" rules, in which an efficient estimator $\hat{P}$ is plugged into the rules derived in Sections 3.2 and 4.2, are not asymptotically efficient-robust if

the key quantities which determine the robust forecast (i.e., $p_L(P)$ and $p_U(P)$ in the binary case) are only directionally differentiable functions of P. This stands in contrast with other asymptotic efficiency results for related problems that depend smoothly on first-stage estimators, including point estimation under partial identification (Song, 2014) and efficient statistical treatment rules under point identification (Hirano and Porter, 2009), for which plug-in rules are efficient. Second, forecasts that are constructed via *bagging* tend to be asymptotically efficient-robust. To construct such forecasts, the posterior distribution for P is replaced by the bootstrap distribution of an efficient estimator of P. The forecast is then chosen to minimize the maximum risk or regret over $\Theta_0(P)$ averaged across the bootstrap distribution. As discussed in Remark 5.6 below, it can be shown that the bagged forecasts are asymptotically equivalent to the Bayesian robust forecasts even under directional differentiability.

5.1 Limit Experiment

Our approach follows Hirano and Porter (2009) and uses Le Cam's limits of experiments framework. As is standard for treatments of asymptotic efficiency (see, e.g., van der Vaart (2000)), we work with a local reparameterization in which the reduced form parameter is $P_{n,h} = P_0 + n^{-1/2}h$ for P_0 fixed and h ranging over $\mathbb{R}^k$. Let $\overset{P_{n,h}}{\rightsquigarrow}$ and $\overset{P_{n,h}}{\to}$ denote convergence in distribution and in probability under the sequence of measures $\{F_{n,P_{n,h}}\}_{n\geq 1}$. The model for X_n is *locally asymptotically normal* at P_0 if for each $h_0 \in \mathbb{R}^k$, the likelihood ratio processes indexed by any finite subset $H \subset \mathbb{R}^k$ converge weakly to the likelihood ratio in a shifted normal model:

$$\left(\frac{\mathrm{d}F_{n,P_{n,h}}}{\mathrm{d}F_{n,P_{n,h_0}}}\right)_{h\in H} \overset{P_{n,h_0}}{\rightsquigarrow} \left(\exp\left((h-h_0)'Z - \frac{1}{2}(h-h_0)'I_0(h-h_0)\right)\right)_{h\in H} \tag{42}$$

with $Z \sim N(h_0, I_0^{-1})$ and I_0 nonsingular. Let $\mathbb{E}_h$ and $\mathbb{P}_h$ denote expectation and probability with respect to $Z \sim N(h, I_0^{-1})$.

Assumption 5.1.

1. *$\mathcal{P}$ is an open subset of $\mathbb{R}^k$ with $P_0 \in \mathcal{P}$;*
2. *The model for X_n is locally asymptotically normal at each $P_0 \in \mathcal{P}$.*

In the dynamic binary choice example, Assumption 5.1.1. implies we observe all possible realizations of histories $Y_i^T \in \{0,1\}^T$ up to time T with positive probability.

To describe the limit experiment, consider the collection $\mathbb{D}$ of sequences of forecasts $\{d_n\}_{n\geq 1}$ that converge in distribution under $\{F_{n,P_{n,h}}\}_{n\geq 1}$:

$$\mathbb{D} = \left\{\{d_n\}_{n\geq 1} : d_n(X_n) \overset{P_{n,h}}{\rightsquigarrow} Q_{P_0,h} \text{ for all } h \in \mathbb{R}^k \text{ and } P_0 \in \mathcal{P}\right\}, \tag{43}$$

where $Q_{P_0,h}$ denotes a probability measure on $\mathcal{D}$ equipped with its Borel σ-algebra. Assumption 5.1 permits application of an asymptotic representation theorem of van der Vaart (1991). For any $\{d_n\}_{n\geq 1} \in \mathbb{D}$ there exists a function $d^{\infty}_{P_0}(Z,U)$ with $d^{\infty}_{P_0}(Z,U) \sim Q_{P_0,h}$ where $Z \sim N(h, I_0^{-1})$ and $U \sim \text{Uniform}[0,1]$ independently of Z is a randomization term. As $n \to \infty$, the average excess maximum risk and regret of any sequence of forecasts $\{d_n\}_{n\geq 1} \in \mathbb{D}$ converges to a limiting counterpart for its representation $d^{\infty}_{P_0}(Z,U)$ in the limit experiment.

5.2 Asymptotic Efficiency

The robust forecasts derived in Sections 3.2 and 4.2 are *oracle* forecasts in the sense that they were obtained under the assumption of knowledge of the true set Θ_0. To distinguish the oracle forecasts from the data-dependent forecasts $d_{mm}(X_n)$ and $d_{mmr}(X_n)$, in what follows we use the notation $d^o_{mm}(P)$ and $d^o_{mmr}(P)$ to denote the oracle forecasts when P is the true reduced-form parameter. To facilitate the asymptotic calculations, we evaluate forecasts by their excess maximum risk or regret relative to the oracle. The *excess maximum risk* of $d_n(X_n)$ is

$$\Delta\mathcal{R}_{mm}(d_n; P, X_n) = \sup_{\theta\in\Theta_0(P)} \mathbb{E}_\theta[\ell(Y, d_n(X_n))] - \sup_{\theta\in\Theta_0(P)} \mathbb{E}_\theta[\ell(Y, d^o_{mm}(P))]\,.$$

Integration over (P, X_n) leads to the *integrated excess maximum risk*

$$\Delta\mathcal{B}^n_{mm}(d_n;\pi) = \int\int \sqrt{n}\Delta\mathcal{R}_{mm}(d_n; P, X_n)\,\mathrm{d}\Pi_n(P|X_n)\,\mathrm{d}F_n(X_n).$$

We standardize by $\sqrt{n}$ and recenter at the maximum risk of the oracle to ensure that $\Delta\mathcal{B}^n_{mm}(d_n;\pi)$ converges to a finite but potentially non-zero limit, though this does not change the ranking of forecasts. Therefore, the Bayes robust forecast under minimax risk also minimizes $\Delta\mathcal{B}^n_{mm}(d_n;\pi)$. Excess maximum regret $\Delta\mathcal{R}_{mmr}$ and integrated excess maximum regret $\Delta\mathcal{B}^n_{mmr}(d_n;\pi)$ are defined similarly, replacing risk in the above display with regret.

To derive the asymptotic counterparts, we begin by calculating a frequentist risk that averages over the data X_n conditional on P and then integrates over P using the prior π. To express the frequentist risk, let $\mathbb{E}_{P_{n,h}}$ denote the expectation with respect to $X_n \sim F_{n,P_{n,h}}$. Using this notation and conducting the change-of-variables from $P_{n,h}$ to h, the *frequentist excess maximum risk* can be expressed as

$$\Delta\mathcal{B}^n_{mm}(d_n; P_0, \pi) = \int \mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}_{mm}\left(d_n, P_{n,h}; X_n\right)\right]\,\pi\left(P_{n,h}\right)\,\mathrm{d}h\,. \tag{44}$$

Here we dropped the Jacobian term that arises from the change-of-variables because it simply scales the average excess maximum risk by a power of n without changing the ranking of forecasts. We

include P_0 in the conditioning set to indicate that the calculations are done locally around P_0. The regret $\Delta\mathcal{B}^n_{mmr}(d_n; P_0, \pi)$ can be expressed in a similar manner. *asymptotically efficient-robust forecasts* are those that minimize $\lim_{n\to\infty} \Delta\mathcal{B}^n_{mm}(d_n; P_0, \pi)$ and $\lim_{n\to\infty} \Delta\mathcal{B}^n_{mmr}(d_n; P_0, \pi)$. Under conditions permitting the interchange of limits and integration, we obtain

$$\lim_{n\to\infty} \Delta\mathcal{B}^n_{mm}(d_n; P_0, \pi) = \pi(P_0) \int \left(\lim_{n\to\infty} \mathbb{E}_{P_{n,h}} \left[\sqrt{n} \Delta\mathcal{R}_{mm} \left(d_n, P_{n,h}; X_n \right) \right] \right) \mathrm{d}h \,,$$

and similarly for regret. The limit in parentheses will depend on the sequence of forecasts $\{d_n\}_{n\geq 1}$ through its representation in the limit experiment. Also note that the asymptotic ranking of forecasts does not depend on π.

The following example illustrates the normalization of the excess maximum risk and the use of the local reparameterization.

Example 7: Local parameters and frequentist excess maximum risk. Suppose that $\mathcal{P} = (0, 1)$, $p_L(P) = P$, and

$$p_U(P) = \begin{bmatrix} \frac{1}{2} & P < \frac{1}{2} \,, \\ (2P - \frac{1}{2}) \wedge 1 & P \geq \frac{1}{2} \,, \end{bmatrix}$$

For a binary loss function with $a_{01} = a_{10} = 1$, the robust forecast takes the form

$$d^o_{mm}(P) = \mathbb{I}[1 \leq p_L(P) + p_U(P)] = \mathbb{I}[P \geq \tfrac{1}{2}]. \tag{45}$$

If the true P is bounded away from $\frac{1}{2}$, then eventually we will learn whether it is less or greater than $\frac{1}{2}$ and make the optimal decision. The most challenging case is when P is very close to $\frac{1}{2}$. Thus, we center the local reparameterization at $P_0 = \frac{1}{2}$. Suppose that under $P_{n,h}$ the frequentist sampling distribution and the Bayesian posterior for P is

$$\hat{P}|P_{n,h} \sim N(P_{n,h}, n^{-1}) \,, \quad P|X_n \sim N(\hat{P}, n^{-1}) \,.$$

The posterior is obtained under a uniform prior on $\mathcal{P}$ when n is large enough so that the truncation effect of the prior at the boundary of $(0, 1)$ is negligible. Under the local reparameterization, the sampling distribution of $\hat{h} = \sqrt{n}(\hat{P} - P_0)$ and the posterior distribution for $h = \sqrt{n}(P - P_0)$ is

$$\hat{h}|(P_0, h_0) \sim N(h_0, 1) \,, \quad h|(X_n, P_0) \sim N(\hat{h}, 1) \,.$$

In this example $\hat{P}$ (equivalently $\hat{h}$) is a sufficient statistic. For any decision $d_n(\hat{h})$, we obtain

$$\sup_{\theta\in\Theta_0(P_{n,h_0})} \mathbb{E}_\theta[\ell(Y, d_n(\hat{h}))] = \begin{cases} 1 - p_L(P_{n,h_0}) & \text{if } d_n(\hat{h}) = 1\,, \\ p_U(P_{n,h_0}) & \text{if } d_n(\hat{h}) = 0\,. \end{cases}$$

Here $1 - p_L(P_{n,h_0}) = 1/2 - n^{-1/2}h_0$ is linear in h_0 whereas

$$p_U(P_{n,h_0}) = \begin{cases} \frac{1}{2} + 2n^{-1/2}h_0 & \text{if } h_0 \geq 0 \\ \frac{1}{2} & \text{if } h_0 < 0\,. \end{cases}$$

Using straightforward algebra it can be shown that

$$\mathbb{E}_{P_{n,h_0}}[\sqrt{n}\Delta\mathcal{R}_{mm}(d_n(\hat{h}), P_{n,h_0}; \hat{h})] = \begin{cases} 3h_0\mathbb{P}_{n,h_0}\big[d_n(\hat{h}) = 0\big] & \text{if } h_0 \geq 0\,, \\ -h_0\mathbb{P}_{n,h_0}\big[d_n(\hat{h}) = 1\big] & \text{if } h_0 < 0\,, \end{cases} \tag{46}$$

where $\mathbb{P}_{n,h_0}$ denotes probabilities under $F_{n,P_{n,h_0}}$. It follows that for any sequence $\{d_n\} \in \mathbb{D}$,

$$\lim_{n\to\infty} \mathbb{E}_{P_{n,h_0}}[\sqrt{n}\Delta\mathcal{R}_{mm}(d_n(\hat{h}), P_{n,h_0}; \hat{h})] = \begin{cases} 3h_0\mathbb{P}_{h_0}\big[d^\infty_{P_0}(Z) = 0\big] & \text{if } h_0 \geq 0\,, \\ -h_0\mathbb{P}_{h_0}\big[d^\infty_{P_0}(Z) = 1\big] & \text{if } h_0 < 0\,, \end{cases}$$

where $Z \sim N(h_0, 1)$ under $\mathbb{P}_{h_0}$.[9] The formula shows that the $\sqrt{n}$ standardization leads to a well-defined non-trivial limit of the frequentist excess maximum risk. □

5.3 Binary forecasts

In this section we show that the efficient robust binary forecasts that were derived in Section 3.3 are optimal. For brevity, we focus on discrete forecasts with $\mathcal{D} = \{0, 1\}$ under binary or classification loss. This class of forecasts is also relevant for its connections with statistical treatment rules.

Say $f : \mathcal{P} \to \mathbb{R}^d$ is *directionally differentiable* at P_0 if the limit

$$\lim_{t\downarrow 0} \frac{f(P_0 + th) - f(P_0)}{t} =: \dot{f}_{P_0}[h]$$

exists for every $h \in \mathbb{R}^k$, in which case $\dot{f}_{P_0}[\cdot]$ is its directional derivative. Note that $\dot{f}_{P_0}[\cdot]$ will be positively homogeneous of degree one but not necessarily linear. If $\dot{f}_{P_0}[h]$ is linear in h we say that f is *fully* differentiable at P_0. We say that the posterior Π_n is *consistent* if $\Pi_n(P \in N|X_n) \xrightarrow{P_0} 1$ for every neighborhood N containing P_0 for each $P_0 \in \mathcal{P}$. Recall that $Z \sim N(h, I_0^{-1})$ in the limit experiment. Let $\mathbb{P}_h$ denote probability statements with respect to Z. Moreover, let $Z^* \sim N(0, I_0^{-1})$ independently of Z and $\mathbb{E}^*$ denote expectation with respect to Z^*.

[9] Here it is without loss of generality to write $d^\infty_{P_0}$ as a function of Z only; see the discussion in Appendix C.2.

Assumption 5.2.

1. (a) *The functions p_U and p_L are everywhere continuous and everywhere directionally differentiable;*
   (b) *The function $x \mapsto \mathbb{P}_h\left(\mathbb{E}^*[\dot{p}_{L,P_0}[Z^* + Z] + \dot{p}_{U,P_0}[Z^* + Z]|Z] \leq x\right)$ is continuous at $x = 0$ for each $h \in \mathbb{R}^k$ and $P_0 \in \mathcal{P}$ with $a_{01}p_L(P_0) + a_{10}p_U(P_0) = a_{01}$ and $p_U(P_0) > \frac{a_{01}}{a_{01}+a_{10}}$;*
2. (a) *The posterior for P is consistent;*
   (b) *For any neighborhood N of P_0 there is $\gamma > \frac{1}{2}$ such that $n^\gamma \Pi_n(P \notin N|X_n) \overset{P_0}{\to} 0$;*
3. (a) *At any $P_0 \in \mathcal{P}$ with $a_{01}p_L(P_0) + a_{10}p_U(P_0) = a_{01}$, for any Borel set A,*

$$\lim_{n\to\infty} F_{n,P_{n,h}}\left(\int \sqrt{n}(f(P) - f(P_0))\,\mathrm{d}\Pi_n(P|X_n) \in A\right) = \mathbb{P}_h\left(\mathbb{E}^*[\dot{f}_{P_0}[Z^* + Z]|Z] \in A\right)$$

   with $f = (p_L, p_U)$;
   (b) *Similarly, for any Borel set A,*

$$\lim_{n\to\infty} F_{n,P_{n,h}}\left(\int \sqrt{n}(f(P) - f(P_0))_+\,\mathrm{d}\Pi_n(P|X_n) \in A\right) = \mathbb{P}_h\left(\mathbb{E}^*[(\dot{f}_{P_0}[Z^* + Z])_+|Z] \in A\right)$$

   with $f = (p_L, p_U)$, where $(\cdot)_+$ is applied element-wise.

The directional differentiability of Assumption 5.2.1 was built into the functional form of $p_U(\cdot)$ in Example 7. Appendix A shows that in a broad class of problems the extreme probabilities $p_L(P)$ and $p_U(P)$ can be expressed as min-max or max-min problems, where the outer optimization is over homogeneous parameters and the inner optimization is a linear or convex program. It follows that $p_L(\cdot)$ and $p_U(\cdot)$ are typically only directionally, rather than fully, differentiable functions.[10] Directional differentiability can also be a feature of models defined via moment inequalities (cf. Example 5).

Assumption 5.2.2(a) holds under standard regularity conditions (see, e.g., Chapter 10.4 of van der Vaart (2000)). For Assumption 5.2.2(b), note that $\Pi_n(P \notin N|X_n)$ typically converge to zero exponentially. For instance, in a normal means model with $\bar{X}_n \sim N(P_{n,h}, (nI_0)^{-1})$ and a flat prior on P, we have $\Pi_n(P \notin N|X_n) = O(e^{-cn})$ for some $c > 0$.[11] More generally, the classical posterior consistency results of Schwartz (1965) establish exponential convergence rates.

Assumption 5.2.3 is simply assuming that a δ-method applies for the posterior distribution of

[10] See, e.g., Theorem 3.1 of Greenberg (1997) for directional differentiability of the value of linear programs, Chapter 4.3 of Bonnans and Shapiro (2000) for directional differentiability of the value of convex programs, and Milgrom and Segal (2002) and Shapiro (2008) for directional differentiability of min-max problems.

[11] Note $P|X_n \sim N(\bar{X}_n, (nI_0)^{-1})$ under a flat prior. Choose $\varepsilon > 0$ so that $B_{2\varepsilon}(P_0) \subset N$. Then $P_0(\bar{X}_n \in B_\varepsilon(P_0)) \to 1$ and whenever $\bar{X}_n \in B_\varepsilon(P_0)$, we have $\Pi_n(P \notin N|X_n) \leq \Pi_n(|P - \bar{X}_n| > \varepsilon|X_n) = O(e^{-cn})$ for any $c < \frac{\varepsilon^2}{2}\lambda_{\min}(I_0)$.

directionally differentiable functionals of P.[12] For a heuristic justification for Assumption 5.2.3(a), consider a normal means model with $\bar{X}_n \sim N(P_{n,h}, (nI_0)^{-1})$. Under a flat prior for P, we have $P|X_n \sim N(\bar{X}_n, (nI_0)^{-1})$. For a directionally differentiable function f:

$$\begin{aligned}
&F_{n,P_{n,h}}\left(\int \sqrt{n}(f(P)-f(P_0))\,\mathrm{d}\Pi_n(P|X_n)\in A\right)\\
&=F_{n,P_{n,h}}\left(\int \sqrt{n}(f(P)-f(P_0))\frac{e^{-\frac{1}{2}(P-\bar{X}_n)'(nI_0)(P-\bar{X}_n)}}{\sqrt{|2\pi(nI_0)^{-1}|}}\,\mathrm{d}P\in A\right)\\
&\approx F_{n,P_{n,h}}\left(\int \dot{f}_{P_0}[\sqrt{n}(P-P_0)]\frac{e^{-\frac{1}{2}(P-\bar{X}_n)'(nI_0)(P-\bar{X}_n)}}{\sqrt{|2\pi(nI_0)^{-1}|}}\,\mathrm{d}P\in A\right)\\
&=F_{n,P_{n,h}}\left(\int \dot{f}_{P_0}[\kappa+\sqrt{n}(\bar{X}_n-P_0)]\frac{e^{-\frac{1}{2}\kappa'(I_0)\kappa}}{\sqrt{|2\pi(I_0)^{-1}|}}\,\mathrm{d}\kappa\in A\right),
\end{aligned}$$

with the final line is by the change of variables $\kappa=\sqrt{n}(P-\bar{X}_n)$. The last integral can be rewritten $\mathbb{E}^*[\dot{f}_{P_0}[Z^*+Z]|Z]$ where $Z^*|Z\sim N(0,I_0^{-1})$ with $Z=\sqrt{n}(\bar{X}_n-P_0)\sim N(h,I_0^{-1})$ under $F_{n,P_{n,h}}$. A similar argument provides a heuristic justification for Assumption 5.2.3(b). In the presentation of the asymptotic efficiency results for the binary forecasts we rely on the following definition:

Definition 5.3. *Given a sequence of forecasts $\{d_n\}_{n\geq 1}\in\mathbb{D}$, we say that d_n is* asymptotically equivalent *to $d_{b,mm}$ if $d_n(X_n)$ and $d_{b,mm}(X_n)$ have the same asymptotic distribution under the sequence of measures $\{F_{n,P_{n,h}}\}_{n\geq 1}$ for all $P_0\in\mathcal{P}$ and $h\in\mathbb{R}^k$.*

Let $\Delta\mathcal{B}^n_{b,mm}(\,\cdot\,;P_0,\pi)$ and $\Delta\mathcal{B}^n_{b,mmr}(\,\cdot\,;P_0,\pi)$ denote integrated excess maximum risk and regret (see display (44)) for binary loss ℓ_b from (8). We require forecasts to satisfy an additional technical condition, namely condition (A.10) in the Appendix, which permits the interchange of limits and integration. This condition can be verified under more primitive conditions (see Remark C.6). Let $\mathbb{D}$ denote the set of all sequences of $\{0,1\}$-valued forecasts that converge in the sense of (43). Theorem 5.4 states that forecasts that are asymptotically equivalent to the Bayes forecasts are asymptotically optimal. A proof is provided in Appendix C.

Theorem 5.4. *(i) Let Assumption 5.1 and parts (a) of Assumption 5.2 hold and let $\tilde{d}_n$ be asymptotically equivalent to $d_{b,mm}$ and satisfy condition (A.10). Then: for all $P_0\in\mathcal{P}$,*

$$\lim_{n\to\infty}\Delta\mathcal{B}^n_{b,mm}(\tilde{d}_n;P_0,\pi)=\inf_{\{d_n\}\in\mathbb{D}}\liminf_{n\to\infty}\Delta\mathcal{B}^n_{b,mm}(d_n;P_0,\pi)\,.$$

(ii) Let Assumptions 5.1 and 5.2 hold and let $\tilde{d}_n$ be asymptotically equivalent to $d_{b,mmr}$ and satisfy

[12] See, e.g., Kitagawa, Montiel Olea, Payne, and Velez (2020) for a formal justification. This may require strengthening our definition of directional differentiability to Hadamard directional differentiability.

condition (A.10). Then: for all $P_0 \in \mathcal{P}$,

$$\lim_{n\to\infty} \Delta\mathcal{B}^n_{b,mmr}(\tilde{d}_n; P_0, \pi) = \inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mmr}(d_n; P_0, \pi)\,.$$

Remark 5.5. The asymptotic efficiency extends to Bayes forecasts derived under priors that differ from the "objective" prior that is used to compute the integrated risk but assign positive density in the neighborhood of P_0. In large samples, the posterior is dominated by the likelihood function and the shape of the prior density does not affect the asymptotic form of the posterior distribution. The optimality result also extends to forecasts derived under a misspecified likelihood function, as long as this likelihood function leads to a large-sample posterior that reproduces the asymptotic form of the posterior under the "true" likelihood function. □

Remark 5.6. Given the asymptotic equivalence of posterior distributions of parameters and the bootstrap distributions of their efficient estimators,[13] bagged estimators that replace posterior averaging with averaging across the bootstrap distribution of an efficient estimator of P will yield asymptotically efficient forecasts under suitable modification of the above regularity conditions. □

Remark 5.7. The optimal forecasting problem with $\mathcal{D} = \{0, 1\}$ has a similar form to the optimal treatment problem studied by Hirano and Porter (2009) and our proofs follow similar arguments. In their setting, the oracle treatment rule is of the form $\mathbb{I}[g(P_0) \geq 0]$ where g is (fully) differentiable. Their asymptotically efficient rule replaces $g(P_0)$ by $g(\hat{P})$ where $\hat{P}$ is an efficient estimator of P_0. When p_L and p_U are directionally, rather than fully, differentiable, the optimal forecasts that we derive are of a different form from plugging $\hat{P}$ into the oracle rules. This difference arises because

$$\int \dot{f}_{P_0}[\sqrt{n}(P - P_0)]\,\mathrm{d}\Pi_n(P|X_n) \neq \dot{f}_{P_0}\left[\int \sqrt{n}(P - P_0)\,\mathrm{d}\Pi_n(P|X_n)\right]$$

under directional differentiability. If p_L and p_U are fully differentiable then both sides of the above display are equal and plugging in $\hat{P}$ into the oracle rule is asymptotically efficient. □

As we formalize in Proposition 5.8 below, asymptotic equivalence to $d_{b,mm}$ (respectively, $d_{b,mmr}$) is a *necessary* condition for a forecast $\tilde{d}_n$ to be asymptotically efficient-robust under minimax risk (respectively, regret) under a side condition ensuring that ties occur with probability zero.

In view of Definition 5.3, we say that asymptotic equivalence fails at P_0 if there exists some $h_* \in \mathbb{R}^k$ for which $d_n(X_n)$ and $d_{b,mm}(X_n)$ have *different* asymptotic distributions under the sequence of measures $\{F_{n,P_{n,h_*}}\}_{n\geq 1}$ with $P_{n,h_*} = P_0 + n^{-1/2}h_*$. A proof for the following proposition is provided in Appendix C.

[13] This equivalence carries over to directionally differentiable function (see, e.g., Kitagawa et al. (2020)).

Proposition 5.8. *(i) Let Assumption 5.1 and parts (a) of Assumption 5.2 hold, let $d_{b,mm}$ satisfy condition (A.10), and let $\{\tilde{d}_n\}_{n\geq 1} \in \mathbb{D}$ be a sequence of forecasts for which $\tilde{d}_n$ and $d_{b,mm}$ are* not *asymptotically equivalent. Then for any P_0 at which asymptotic equivalence of $\tilde{d}_n$ and $d_{b,mm}$ fails:*

$$\liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(\tilde{d}_n; P_0, \pi) > \inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(d_n; P_0, \pi)$$

provided either (a) or (b) holds:
(a) $a_{01}p_L(P_0) + a_{10}p_U(P_0) \neq a_{01}$,
(b) $a_{01}p_L(P_0) + a_{10}p_U(P_0) = a_{01}$ and $\mathbb{E}^*\big[a_{01}\dot{p}_{L,P_0}[Z^* + Z] + a_{10}\dot{p}_{U,P_0}[Z^* + Z] \,|\, Z\big] \neq 0$ a.e.

(ii) Let Assumptions 5.1 and 5.2 hold, let $d_{b,mmr}$ satisfy condition (A.10), and let $\{\tilde{d}_n\}_{n\geq 1} \in \mathbb{D}$ be a sequence of forecasts for which $\tilde{d}_n$ and $d_{b,mmr}$ are not *asymptotically equivalent. Then for any P_0 at which asymptotic equivalence of $\tilde{d}_n$ and $d_{b,mmr}$ fails:*

$$\liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mmr}(\tilde{d}_n; P_0, \pi) > \inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mmr}(d_n; P_0, \pi)$$

provided either (a), (b), or (c) holds with $a = \frac{a_{01}}{a_{01}+a_{10}}$:
(a) $p_L(P_0) + p_U(P_0) \neq 2a$,
(b) $p_L(P_0) + p_U(P_0) = 2a$, $p_U(P_0) > p_L(P_0)$, and $\mathbb{E}^*\big[\dot{p}_{L,P_0}[Z^* + Z] + \dot{p}_{U,P_0}[Z^* + Z] \,|\, Z\big] \neq 0$ a.e.,
(c) $p_L(P_0) = p_U(P_0) = a$ and $\mathbb{E}^*\big[(\dot{p}_{L,P_0}[Z^* + Z])_- + (\dot{p}_{U,P_0}[Z^* + Z])_+ \,|\, Z\big] \neq 0$ a.e.

Example 7: (continued). The oracle forecast under the minimax **risk** criterion was given in display (45). In order to compute the Bayesian robust forecast we need to evaluate

$$\begin{aligned}
&\int (p_L(P) + p_U(P))\, \mathrm{d}\Pi_n(P|X_n) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \left[1/2 + n^{-1/2}h + 1/2\right] \exp\left\{-\frac{1}{2}(h-\hat{h})^2\right\} \mathrm{d}h \\
&\quad + \frac{1}{\sqrt{2\pi}} \int_{0}^{+\infty} \left[1/2 + n^{-1/2}h + 1/2 + 2n^{-1/2}h\right] \exp\left\{-\frac{1}{2}(h-\hat{h})^2\right\} \mathrm{d}h \\
&= 1 + n^{-1/2}\left[\hat{h} + 2\Phi_N(\hat{h})\hat{h} + 2\phi_N(\hat{h})\right],
\end{aligned}$$

where $\hat{h} = \sqrt{n}(\hat{P} - P_0)$ with $P_0 = \frac{1}{2}$ and Φ_N and ϕ_N denote the standard normal cdf and pdf. As $\hat{P}$ is a sufficient statistic, the forecasts only depend on the data X_n though $\hat{P}$ or, equivalently, through $\hat{h}$. We may deduce that

$$d_{b,mm}(\hat{h}) = \mathbb{I}\left[\hat{h} \geq -\frac{2\phi_N(\hat{h})}{1 + 2\Phi_N(\hat{h})}\right]. \tag{47}$$

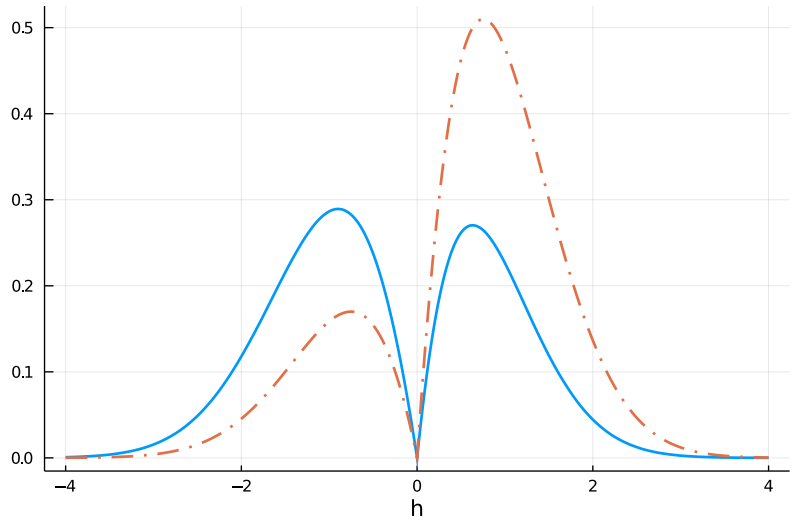


Figure 2: Frequentist excess maximum **risk** in the limit experiment of the efficient robust forecast $d_{b,mm}$ (solid line) and the forecast d_{mm}^{plug} based on plugging in an efficient estimator of P (dot-dashed line) as a function of the location parameter h_0 for the Example 7.

Note that the term on the right-hand side of the inequality in the indicator function is always negative. The plug-in forecast is obtained by replacing the unknown P by $\hat{P}$, leading to $d_{b,mm}^{plug}(\hat{P}) = \mathbb{I}[p_L(\hat{P}) + p_U(\hat{P}) \geq 1]$. In the present example, the plug-in forecast can be expressed equivalently in terms of $\hat{h}$:

$$d_{b,mm}^{plug}(\hat{h}) = \mathbb{I}\left[\hat{h} \geq 0\right]. \tag{48}$$

The plug-in forecast is not asymptotically equivalent to the Bayesian robust forecast. In particular, the Bayesian robust forecast predicts $Y = 1$ more aggressively than the plug-in forecast. It can be verified by direct calculation in this example that this more aggressive forecast is asymptotically efficient-robust. It also follows from Proposition 5.8 that the plug-in forecast is not asymptotically efficient-robust. This is seen by noting that $a_{01} = a_{10} = 1$, $p_L(P_0) + p_U(P_0) = 1$, $\dot{p}_{L,P_0}[h] = h$, $\dot{p}_{U,P_0}[h] = 2(h)_+$, and so $\mathbb{E}^*\left[a_{01}\dot{p}_{L,P_0}[Z^*+Z]+a_{10}\dot{p}_{U,P_0}[Z^*+Z] \,\middle|\, Z\right]$ reduces to $Z+2\mathbb{E}^*[(Z^*+Z)_+ \,|\, Z]$ which is nonzero almost everywhere.

To quantify the inefficiency of $d_{b,mm}^{plug}(\hat{h})$ relative to $d_{b,mm}(\hat{h})$, straightforward algebraic manipulations using (46) allow us to derive formulas for the frequentist excess maximum risk of $d_{b,mm}(\hat{h})$ and $d_{b,mm}^{plug}(\hat{h})$ as a function of h_0. The results are plotted in Figure 2. The plug-in forecast is inferior to the Bayesian robust forecast from an integrated risk perspective: the area under the curve corresponding to $d_{b,mm}$ is around 20% smaller than that under the curve corresponding to the plug-in forecast. While $d_{b,mm}$ was designed to be optimal from an integrated risk perspective, it also dominates the plug-in forecast from a minimax perspective: the maximum excess maximum risk of the plug-in forecast in the limit experiment is around 75% larger than that of $d_{b,mm}$.

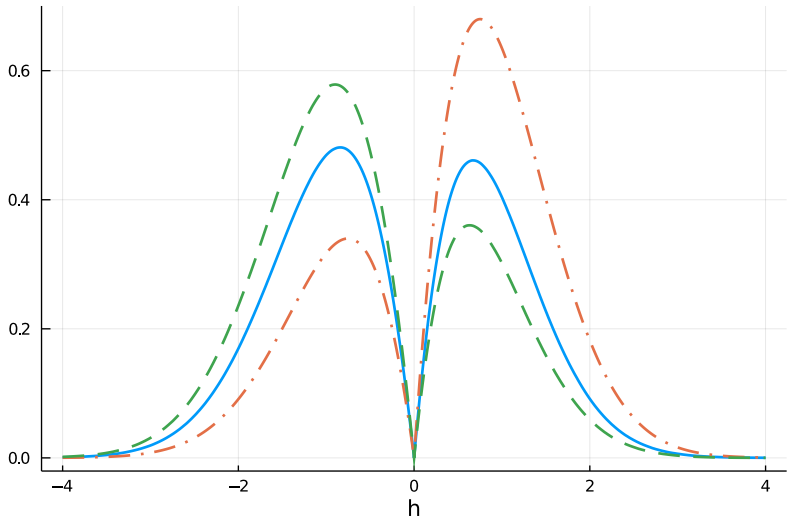


Figure 3: Frequentist excess maximum **regret** in the limit experiment of the efficient robust forecast $d_{b,mmr}$ (solid line), the forecast d^{plug}_{mmr} based on plugging an efficient estimator of P (dot-dashed line), and the forecast $d_{b,mm}$ (dashed line) as a function of the location parameter h_0 for the Example 7.

Similar calculations can be made under the **regret** criterion. The oracle forecast is of the form $d^o_{b,mmr}(P) = \mathbb{I}[(\frac{1}{2} - p_L(P))_+ \leq (p_U(P) - \frac{1}{2})_+]$. Similar calculations as for the risk criterion can be used to obtain a formula for the Bayesian robust forecast. In turns out that the plug-in forecast remains unchanged. Frequentist excess maximum regrets as a function of h_0 are plotted in Figure 3. Again, the plug-in forecast is not asymptotically efficient-robust and dominated by the Bayesian efficient robust forecast once we average across h_0. Its integrated excess maximum regret is around 8% smaller and maximum excess maximum regret is around 41% smaller. Also shown is the excess maximum regret of a forecast which plugs the posterior means of $p_L(P)$ and $p_U(P)$ into the oracle: $d^\dagger(X_n) = \mathbb{I}[(\frac{1}{2} - \int p_L(P)\,\mathrm{d}\Pi_n(P|X_n))_+ \leq (\int p_U(P)\,\mathrm{d}\Pi_n(P|X_n) - \frac{1}{2})_+]$. This forecast is equivalent to the minimax forecast $d_{b,mm}$ and is therefore optimal for minimizing integrated excess maximum risk but not necessarily integrated excess maximum regret. Figure 3 shows that $d_{b,mmr}$ also dominates $d^\dagger$ in terms of both its average (2.5% smaller) and maximum (21% smaller) excess maximum regret in the limit experiment. □

Remark 5.9. Consider the numerical example from Section 3.4. The optimization problems $p_U(P)$ and $p_L(P)$ can be recast as the value of max-min and min-max problems in which the reduced-form parameter P enters the objective function. As is well known (Milgrom and Segal, 2002; Shapiro, 2008), the value of max-min and min-max problems is typically only directionally, rather than fully, differentiable. □

5.4 Multinomial Forecasts

We now turn to extending the asymptotic efficiency result to multinomial forecasts that are asymptotically equivalent to the Bayesian robust forecast from Section 4.3. To do so, we first state some additional regularity conditions.

Assumption 5.10.

1. (a) *The functions* $\underline{p}_0, \ldots, \underline{p}_M$ *are everywhere continuous and everywhere directionally differentiable;*
   (b) *The functions* $\Delta p_0, \ldots, \Delta p_M$ *are everywhere continuous and everywhere directionally differentiable;*
2. *The posterior for* P *is consistent;*
3. (a) *At any* $P_0 \in \mathcal{P}$ *with* $\underline{p}_m(P_0) = \underline{p}_{m'}(P_0)$ *for some* $m' \neq m$ *and* $\underline{p}_m(P_0) \geq \underline{p}_k(P_0)$ *for all* $k \in \{0, \ldots, M\}$*, for any Borel set* A *we have*

$$\lim_{n\to\infty} F_{n,P_{n,h}}\left(\int \sqrt{n}(f(P) - f(P_0))\, \mathrm{d}\Pi_n(P|X_n) \in A\right) = \mathbb{P}_h\left(\mathbb{E}^*[\dot{f}_{P_0}[Z^* + Z]|Z] \in A\right)$$

with $f = (\underline{p}_0, \ldots, \underline{p}_M)$*;*

   (b) *At any* $P_0 \in \mathcal{P}$ *with* $\Delta p_m(P_0) = \Delta p_{m'}(P_0)$ *for some* $m' \neq m$ *and* $\Delta p_m(P_0) \leq \Delta p_k(P_0)$ *for all* $k \in \{0, \ldots, M\}$*, for any Borel set* A *we have*

$$\lim_{n\to\infty} F_{n,P_{n,h}}\left(\int \sqrt{n}(f(P) - f(P_0))\, \mathrm{d}\Pi_n(P|X_n) \in A\right) = \mathbb{P}_h\left(\mathbb{E}^*[\dot{f}_{P_0}[Z^* + Z]|Z] \in A\right)$$

with $f = (\Delta p_0, \ldots, \Delta p_M)$*;*

Assumption 5.10 is similar to Assumption 5.2. In particular, a heuristic justification for Assumption 5.10.3 follows similar reasoning to that presented earlier for Assumption 5.2.3.

We now present the asymptotic efficiency results for multinomial forecasts. Let $\Delta\mathcal{B}^n_{c,mm}(\,\cdot\,; P_0, \pi)$ and $\Delta\mathcal{B}^n_{c,mmr}(\,\cdot\,; P_0, \pi)$ denote integrated excess maximum risk and regret (see display (44)) for classification loss ℓ_c from (33). Also let $\mathbb{D}$ denote the set of all sequences of $\{0, \ldots, M\}$-valued forecasts that converge in the sense of (43).

Theorem 5.11. *(i) Let Assumption 5.1 and Assumption 5.10.1(a), 5.10.2, and 5.10.3(a) hold and let* $\tilde{d}_n$ *be asymptotically equivalent to* $d_{c,mm}$ *and satisfy condition (A.10). Then: for all* $P_0 \in \mathcal{P}$*,*

$$\lim_{n\to\infty} \Delta\mathcal{B}^n_{c,mm}(\tilde{d}_n; P_0, \pi) = \inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n_{c,mm}(d_n; P_0, \pi)\,.$$

(ii) Let Assumption 5.1 and Assumption 5.10.1(b), 5.10.2, and 5.10.3(b) hold and let $\tilde{d}_n$ be asymptotically equivalent to $d_{c,mmr}$ and satisfy condition (A.10). Then: for all $P_0 \in \mathcal{P}$,

$$\lim_{n\to\infty} \Delta\mathcal{B}^n_{c,mmr}(\tilde{d}_n; P_0, \pi) = \inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n_{c,mmr}(d_n; P_0, \pi)\,.$$

As with Remark 5.6, bagged forecasts in which the posterior distribution is replaced with the bootstrap distribution of an efficient estimator of P can be shown to be asymptotically efficient-robust under a suitable modification of the regularity conditions. As with Proposition 5.8, it is possible to show that forecasts that are not asymptotically equivalent to the $d_{c,mm}$ and $d_{c,mmr}$ are not asymptotically efficient-robust under side conditions ruling out ties.

6 Conclusion

In this paper we proposed use of robust forecasts that are obtained by solving a minimax risk or minimax regret problem to deal with uncertainty about the forecast distribution. We also derived asymptotically efficient-robust forecasts that deal with the estimation of the set of forecast distributions. In addition to being useful for forecasting binary and multinomial outcomes, these methods have wide applicability in environments in which a forecaster is concerned about structural breaks, model misspecification, or a policy maker has to make treatment assignments.

References

Baltagi, B. H. (2008). Forecasting with panel data. *Journal of Forecasting 27*(2), 153–173.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York.

Bonhomme, S. and M. Weidner (2019). Minimizing sensitivity to model misspecification. Manuscript, University of Chicago.

Bonnans, J. F. and A. Shapiro (2000). *Perturbation Analysis of Optimization Problems*. Springer.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*(2), 123–140.

Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In J. J. Heckman and B. S. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*, Econometric Society Monographs, pp. 3–38. Cambridge University Press.

Chamberlain, G. (2000). Econometric applications of maxmin expected utility. *Journal of Applied Econometrics 15*(6), 625–644.

Chamberlain, G. (2001). Minimax estimation and forecasting in a stationary autoregression model. *American Economic Review, Papers & Proceedings 91*(2), 55–59.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica 78*(1), 159–168.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Christensen, T. and B. Connault (2019). Counterfactual sensitivity and robustness. *Manuscript, New York University*.

Ciliberto, F. and E. Tamer (2009). Market structure and multiple equilibria in airline markets. *Econometrica 77*(6), 1791–1828.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological) 20*(2), 215–232.

Csiszár, I. and F. Matúš (2012). Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. *Kybernetika 48*(4), 637–689.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics 125*(1), 141–173.

Elliott, G. and R. P. Lieli (2013). Predicting binary outcomes. *Journal of Econometrics 174*, 15–26.

Elliott, G. and A. Timmermann (2016). *Economic Forecasting*. Princeton University Press, Princeton.

Giacomini, R. and T. Kitagawa (2018). Robust bayesian inference for set-identified models. *Manuscript, University College London*.

Giacomini, R., T. Kitagawa, and H. Uhlig (2019). Estimation under ambiguity. *cemmap working paper No. CWP24/19*.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics 18*(2), 141–153.

Greenberg, H. J. (1997). Linear programming 1: Basic principles. In T. Gal and H. J. Greenberg (Eds.), *Advances in Sensitivity Analysis and Parametic Programming*, pp. 57–100. Springer.

Grieco, P. L. E. (2014). Discrete games with flexible information structures: an application to local grocery markets. *The RAND Journal of Economics 45*(2), 303–340.

Gu, J. and R. Koenker (2016). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics (Forthcoming)*.

Hansen, L. P. and T. J. Sargent (2001). Robust control and model uncertainty. *The American Economic Review 91*(2), 60–66.

Heckman, J. J., J. Smith, and N. Clements (1997). Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts. *The Review of Economic Studies 64*(4), 487–535.

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*(5), 1683–1701.

Hirano, K. and J. H. Wright (2017). Forecasting with model uncertainty: Representations and risk reduction. *Econometrica 85*(2), 617–643.

Honoré, B. E. and E. Kyriazidou (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica 68*(4), 839–874.

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*(5), 611–629.

Jia, P. (2008). What happens when Wal-Mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica 76*(6), 1263–1316.

Khan, S., F. Ouyang, and E. Tamer (2019). Inference on Semiparametric Multinomial Response Models. *Manuscript*.

Kitagawa, T. (2012). Estimation and inference for set-identified parameters using posterior lower probabilities. *Manuscript, University College London*.

Kitagawa, T., J. L. Montiel Olea, J. Payne, and A. Velez (2020). Posterior distribution of nondifferentiable functions. *Journal of Econometrics 217*(1), 161–175.

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

Lahiri, K. and L. Yang (2013). Forecasting binary outcomes. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, pp. 1025–1106. Elsevier, New York.

Liu, L. (2019). Density forecasts in panel data models: A semiparametric bayesian perspective. *Manuscript, Indiana University*.

Liu, L., H. R. Moon, and F. Schorfheide (2018). Forecasting with a panel tobit model. *Manuscript, University of Pennsylvania*.

Liu, L., H. R. Moon, and F. Schorfheide (2020). Forecasting with dynamic panel data models. *Econometrica 88*(1), 171–201.

Magnac, T. (2000). Subsidised training and youth employment: Distinguishing unobserved heterogeneity from state dependence in labour market histories. *The Economic Journal 110*(466), 805–837.

Manski, C. F. (1996). Learning about treatment effects from experiments with random assignment of treatments. *The Journal of Human Resources 31*(4), 709–733.

Manski, C. F. (2000). Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics 95*(2), 415–442.

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

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

Manski, C. F. (2007). Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics 139*(1), 105–115. Endogeneity, instruments and identification.

Manski, C. F. and A. Tetenov (2007). Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *Journal of Statistical Planning and Inference 137*(6), 1998–2010.

Milgrom, P. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *Econometrica 70*(2), 583–601.

Moon, H. R. and F. Schorfheide (2012). Bayesian and frequentist inference in partially identified models. *Econometrica 80*(755–782).

Robbins, H. (1951). Asymptocially subminimax solutions of compound decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Volume I. University of California Press, Berkeley and Los Angeles.

Schwartz, L. (1965). On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 4*(1), 10–26.

Shapiro, A. (2008). Asymptotics of minimax stochastic programs. *Statistics & Probability Letters 78*(2), 150–157.

Song, K. (2014). Point decisions for interval-identified parameters. *Econometric Theory 30*(2), 334–356.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics 166*(1), 157–165.

van der Vaart, A. (1991). An asymptotic representation theorem. *International Statistical Review 59*(1), 97–121.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.

Wald, A. (1950). *Statistical Decision Functions*. John Wiley, New York.

Online Appendix: Robust Forecasting

Timothy Christensen, Hyungsik Roger Moon, and Frank Schorfheide

A Computation

The challenge in implementing the minimax and minimax regret forecasts is to solve the extremum problems p_L and p_U from (15) and (16) in the binary case, or $\underline{p}_m$ and Δp_m from (35) and (39) in the multinomial case.

We show how to compute these quantities in a class of models in which (i) vector of model parameters θ may be partitioned as $\theta = (\phi, \Pi)$, where ϕ is a low-dimensional parameter and Π is a probability measure, and (ii) both the forecast probabilities and restrictions defining set Θ_0 are linear in Π. This nests semiparametric panel data models we study (Examples 1–4) and several other models, such as game-theoretic models (Example 5). In the next subsection, we show how linear programming techniques similar to Honoré and Tamer (2006) may be used when the support of Π is discrete. Subsection A.2 studies the continuous case.

A.1 Computing Extreme Probabilities: the Discrete Case

A.1.1 Binary forecasts

We consider a class of problems where the forecast probabilities and restrictions that define the set Θ_0 are linear in Π, where Π has discrete support. We can identify Π with a vector $\pi \in \Delta^{L-1}$, where L is the number of points of support of Π and $\Delta^{L-1} = \{x \in \mathbb{R}^L_+ : \sum_{i=1}^L x_i = 1\}$. We further assume that we can write the forecast probability as

$$b(\phi)'\pi\,, \tag{A.1}$$

where $b(\phi)$ is a L-vector that may depend on the homogeneous parameters, and the restrictions defining Θ_0 as

$$G(\phi)\pi = r\,, \tag{A.2}$$

where $G(\phi)$ is a $K \times L$ matrix and $r \in \mathbb{R}^K$.

Consider, for example, the semiparametric panel data model (Example 1). In that setting, the low-dimensional parameter ϕ is β, the probability measure Π is the joint distribution $\Pi_{\lambda,y}$ of (λ_i, Y_{i0}), and the parameter space is $\Theta = \{(\beta, \Pi_{\lambda,y})\}$. The identified set is the collection of all

$(\beta, \Pi_{\lambda,y})$ such that the model-implied probabilities of observing each realization of Y_i^T is equal to the population probability $p(y^T)$; see display (2). The model-implied probabilities are given by

$$p(y^T|\beta, \Pi_{\lambda,y}) = \int p(y^T|y_0, \lambda; \beta)\, \mathrm{d}\Pi_{\lambda,y}(\lambda, y_0)\,,$$

with $p(y^T|y_0, \lambda; \beta)$ from expression (31). Because $p(y^T|\beta, \Pi_{\lambda,y}) = p(y^T)$ for any $\theta \in \Theta_0$, the forecast probability given $Y_i^T = y^T$ is

$$\mathbb{P}_\theta(Y_{iT+1} = 1|Y_i^T = y^T) = \frac{\int \Phi(\beta y_{iT} + \lambda) p(y^T|y_0, \lambda; \beta) \mathrm{d}\Pi_{\lambda,y}(\lambda, y_0)}{p(y^T)}\,.$$

Returning to the general case with forecast probabilities as in (A.1) and restrictions defining Θ_0 as in (A.2), we can write p_U as

$$p_U = \sup_\phi \left(\sup_{\pi \in \Delta^{L-1}} b(\phi)'\pi \quad \text{s.t.} \quad G(\phi)\pi = r \right)\,.$$

As we show in the following proposition, the inner optimization over π can be written as a linear program, simplifying computation.

Proposition A.1. *The program*

$$p_\phi = \sup_{\pi \in \Delta^{L-1}} b(\phi)'\pi \quad \textit{s.t.} \quad G(\phi)\pi = r$$

has an equivalent dual formulation

$$p_\phi^* = \inf_{v \in \mathbb{R}^{K+1}} [0_{1\times K}\,,\, 1]\, v \quad \textit{s.t.} \quad A(\phi)v \leq -b(\phi)$$

where $A(\phi) = [G(\phi)' - (1_{L\times 1} \otimes r')\,,\, -1_{L\times 1}]$ *with* $\otimes$ *denoting the Kronecker product.*

In view of Proposition A.1, we may compute p_U by solving

$$p_U = \sup_\phi \left(\inf_{v \in \mathbb{R}^{K+1}} [0_{1\times K}\,,\, 1]\, v \quad \text{s.t.} \quad A(\phi)v \leq -b(\phi) \right)\,. \tag{A.3}$$

If ϕ is not feasible, i.e., if there does not exist $\pi \in \Delta^{L-1}$ solving (A.2), then the inner linear program returns no solution. In this case, we set the value of the inner minimization problem to $-\infty$. The smallest forecast probability p_L is computed similarly:

$$p_L = \inf_\phi \left(\sup_{v \in \mathbb{R}^{K+1}} [0_{1\times K}\,,\, -1]'\, v \quad \text{s.t.} \quad A(\phi)v \leq b(\phi) \right)\,, \tag{A.4}$$

where we set the value of the inner linear program to $+\infty$ if it has no solution.

A.1.2 Multinomial forecasts

In the multinomial case, we consider a setting in which Θ_0 is defined as in display (A.2) for suitable $G(\phi)$ and the forecast probabilities of each of the outcomes $m = 0, 1, \ldots, M$ can be written as

$$b_m(\phi)'\pi$$

for each m.

For minimax forecasts, the lower probabilities $\underline{p}_m$ from (35) are computed analogously to p_U, replacing $b(\phi)$ in (A.4) with $b_m(\phi)$:

$$\underline{p}_m = \inf_{\phi} \left(\sup_{v \in \mathbb{R}^{K+1}} [0_{1\times K}\,,\, -1]'\, v \quad \text{s.t.} \quad A(\phi)v \leq b_m(\phi) \right),$$

for $m = 0, 1, \ldots, M$, where we set the value of the inner linear program to $+\infty$ if it has no solution. For minimax regret forecasts, the terms Δp_m from (39) can be computed analogously to (A.3). To do so, first note that for each $m' = 0, 1, \ldots, M$ we can compute

$$\sup_{\theta \in \Theta_0} \left(\mathbb{P}_\theta(Y = m') - \mathbb{P}_\theta(Y = m) \right)$$

by replacing the term $b(\phi)$ in (A.3) with $b_{m'}(\phi) - b_m(\phi)$. The value Δp_m is then the maximum over all such m':

$$\Delta p_m = \max_{m'} \sup_{\phi} \left(\inf_{v \in \mathbb{R}^{K+1}} [0_{1\times K}\,,\, 1]\, v \quad \text{s.t.} \quad A(\phi)v \leq (b_m(\phi) - b_{m'}(\phi)) \right),$$

where we again set the value of the inner linear program to $-\infty$ if it has no solution.

A.2 Computing Extreme Probabilities: the Continuous Case

A.2.1 Binary forecasts

We first consider a class of problems where the forecast probabilities and restrictions that define Θ_0 are linear in Π, where Π is a probability measure on $(X, \mathcal{X})$ where $\mathcal{X}$ denotes the Borel σ-field on X. We restrict Π to have density with respect to some σ-finite dominating measure ν (e.g. Lebesgue measure) and identify each Π with its density π with respect to ν.[14] We consider a setting where

[14] This nests the previous discrete case by taking X to be the set of L points of discrete support for Π and ν to be counting measure.

forecast probabilities can be written analogously to (A.1) as

$$\int b(x;\phi)\pi(x)\,\mathrm{d}\nu(x)$$

where $b(\,\cdot\,;\phi) : X \to \mathbb{R}$ is a bounded function for each ϕ. We first consider a class of problems in which the set Θ_0 is defined via a moment restriction similar to (A.2), namely

$$\int g(x;\phi)\pi(x)\,\mathrm{d}\nu(x) = r\,,$$

where $g(\,\cdot\,;\phi) : X \to \mathbb{R}^K$ is a vector of moment functions.

The semiparametric panel data model (Example 1) is of this form, where we now relax the assumption of discrete support for (λ, y_0) and allow the joint distribution $\Pi_{\lambda,y}$ to be an arbitrary distribution on $\mathbb{R} \times \{0,1\}$. The dominating measure ν is the product of Lebesgue measure on $\mathbb{R}$ and counting measure on $\{0,1\}$.

Let Π_ϕ denote the set of all densities π with respect to ν, for which $\int g(x;\phi)\pi(x)\,\mathrm{d}\nu(x)$ is finite and $\int \pi(x)\,\mathrm{d}\nu(x) = 1$. We then have

$$\Theta_0 = \left\{(\phi,\pi) : \pi \in \Pi_\phi\,, \int g(x;\phi)\pi(x)\,\mathrm{d}\nu(x) = r\right\}. \tag{A.5}$$

In this setting, we can write p_U as

$$p_U = \sup_\phi \left(\sup_{\pi\in\Pi_\phi} \int b(x;\phi)\pi(x)\,\mathrm{d}\nu(x) \quad \text{s.t.} \quad \int g(x;\phi)\pi(x)\,\mathrm{d}\nu(x) = r \right).$$

The inner optimization over π has a dual program. Although this dual formulation does not simplify computation a great deal, it can be approximated by a more tractable, finite-dimensional convex program. In what follows, let $\mathrm{ri}(A)$ denote the relative interior of a set A. The following proposition collects results from Csiszár and Matúš (2012) (for the dual formulation) and Christensen and Connault (2019) (for the approximation by a finite-dimensional convex program).

Proposition A.2. *If*

$$r \in \mathrm{ri}\left(\left\{\int g(x;\phi)\pi(x)\,\mathrm{d}\nu(x) : \pi \in \Pi_\phi\right\}\right),$$

then the program

$$p_\phi = \sup_{\pi\in\Pi_\phi} \int b(x;\phi)\pi(x)\,\mathrm{d}\nu(x) \quad \textit{s.t.} \quad \int g(x;\phi)\pi(x)\,\mathrm{d}\nu(x) = r$$

has an equivalent dual formulation

$$p_\phi^* = \inf_{\mu:\nu\text{-ess sup}_x(b(x;\phi)+\mu'(g(x;\phi)-r))<+\infty} \left(\nu\text{-ess sup}_x \left(b(x;\phi) + \mu'(g(x;\phi) - r) \right) \right).$$

In addition, if Π^* *has a strictly positive density* $\pi \in \Pi_\phi$ *and* $\mathbb{E}^{\Pi_*}\left[e^{c\|g(X;\phi)\|}\right]$ *is finite for each* $c \geq 0$*, then*

$$p_\phi^* = \lim_{\delta\to\infty} \left(\sup_{\eta\geq 0,\mu} -\eta \log \mathbb{E}^{\Pi_*}\left[e^{-\eta^{-1}(b(X;\phi)+\mu'(g(X;\phi)-r))}\right] - \eta\delta \right),$$

where $\mathbb{E}^{\Pi_*}[\,\cdot\,]$ *denotes expectation is taken under the distribution* Π_**.*

In view of Proposition A.2, we may compute p_U using the approximation

$$p_U \approx \sup_\beta \left(\inf_{\eta\geq 0,\mu} \eta \log \mathbb{E}^{\Pi_*}\left[e^{\eta^{-1}(b(X;\phi)+\mu'(g(X;\phi)-r))}\right] \right) + \eta\delta\,,$$

which is valid for large δ. The lower probability p_L can be computed analogously:

$$p_L \approx \inf_\beta \left(\sup_{\eta\geq 0,\mu} -\eta \log \mathbb{E}^{\Pi_*}\left[e^{-\eta^{-1}(b(X;\phi)+\mu'(g(X;\phi)-r))}\right] \right) - \eta\delta\,. \tag{A.6}$$

Similar techniques may also be used when Θ_0 arises out of robustness concerns; see Example 2. To that end, we can consider a class of models where forecast probabilities and restrictions defining Θ_0 are linear in Π, but where we now restrict Π to the class

$$\Pi_{\phi,\delta} = \{\Pi : K(\Pi\|\Pi_\phi) \leq \delta\}\,,$$

where $\delta \geq 0$ and $K(\Pi\|\Pi_\phi)$ is the Kullback–Leibler divergence between Π and a reference density Π_ϕ. In the context of Example 2, Π_ϕ is a correlated random effects distribution indexed by auxiliary parameters ξ, and $\phi = (\beta, \xi)$. The identified set is now

$$\Theta_0 = \left\{ (\phi, \Pi) : \Pi \in \Pi_{\phi,\delta}\,, \int g(x;\phi)\,\mathrm{d}\Pi(x) = r \right\}. \tag{A.7}$$

With this notion of the identified set, we may apply well known duality methods to compute the extreme probabilities using the dual representations

$$\begin{aligned} p_U &= \sup_\phi \left(\inf_{\eta\geq 0,\mu} \eta \log \mathbb{E}^{\Pi_\phi}\left[e^{\eta^{-1}(b(X;\phi)+\mu'(g(X;\phi)-r))}\right] \right) + \eta\delta\,, \\ p_L &= \inf_\phi \left(\sup_{\eta\geq 0,\mu} -\eta \log \mathbb{E}^{\Pi_\phi}\left[e^{-\eta^{-1}(b(X;\phi)+\mu'(g(X;\phi)-r))}\right] \right) - \eta\delta\,, \end{aligned} \tag{A.8}$$

which are valid whenever $\mathbb{E}^{\Pi_\phi}\left[e^{c\|g(X;\phi)\|}\right]$ is finite for each $c \geq 0$ and each ϕ, and

$$r \in \operatorname{ri}\left(\left\{\int g(x;\phi)\,\mathrm{d}\Pi(x) : \Pi \in \Pi_{\phi,\delta}\right\}\right);$$

see, e.g., Christensen and Connault (2019) for a formal statement. Similar dual representations apply for neighborhoods constrained by other ϕ-divergences.

A.2.2 Multinomial forecasts

Multinomial forecasts can be implemented similarly using the reformulations described above. For minimax forecasts, if the forecast probabilities are each of the form

$$\int b_m(x,\phi)\,\mathrm{d}\Pi(x)$$

for $m = 0, 1, \ldots, M$, then each $\underline{p}_m$ can be computed as in (A.6) or (A.8), replacing b with b_m. For minimax regret forecasts, each Δp_m can be computed as

$$\Delta p_m \approx \max_{m'} \sup_{\beta} \left(\inf_{\eta \geq 0, \mu} \eta \log \mathbb{E}^{\Pi_*}\left[e^{\eta^{-1}(b_{m'}(X;\phi) - b_m(X;\phi) + \mu'(g(X;\phi) - r))}\right]\right) + \eta\delta\,,$$

when Θ_0 is of the form (A.5). A similar computation applies when Θ_0 is of the form (A.7), replacing Π_* with Π_ϕ.

B Further Results on Robust Binary Forecasts

B.1 Equivalence of Minimax forecasts under Quadratic and Logarithmic Loss

Here we show that the minimax forecast under quadratic loss is also minimax under logarithmic loss. We first rule out a few pathological cases. Suppose the econometrician chooses $d = 0$. If $p_U > 0$ then the the maximum risk is $+\infty$, which is obtained by the maximizing agent choosing any $\theta \in \Theta_0$ with $\mathbb{P}_\theta(Y = 1) > 0$. Thus, it is only optimal to choose $d = 0$ when $p_U = 0$, in which case $\mathbb{P}_\theta(Y = 1) = 0$ for all $\theta \in \Theta_0$. A parallel argument shows it is only optimal to choose $d = 1$ when $p_L = 1$. More generally, if $p_L = p_U$ then it is optimal to choose d to be their common value. Now suppose that $p_L < p_U$. Problem (4) becomes

$$\begin{aligned}\inf_{d\in\mathcal{D}} \sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\ell_p(Y,d)] &= \inf_{d\in[0,1]} \sup_{p\in[p_L,p_U]} -p\log d - (1-p)\log(1-d) \\ &= \sup_{p\in[p_L,p_U]} \inf_{d\in[0,1]} -p\log d - (1-p)\log(1-d)\,,\end{aligned}$$

where the first equality is because for any $d \in [0,1]$, the maximum risk is obtained at either p_L or p_U, and the second equality is by the minimax theorem. The inner minimum is achieved at $d = p$, and the outer maximum is achieved by taking $p \in [p_L, p_U]$ to be as close to $\frac{1}{2}$ as possible.

B.2 Equivalence of Robust Binary Forecasts under Classification Loss

We now show that the minimax and minimax regret forecasts are identical under classification loss. First suppose $p_L > \frac{1}{2}$. In this case, the θ-optimal decision is $d^*_{b,\theta} = 1$ for all $\theta \in \Theta_0$ and so $d_{b,mmr} = d_{b,mm} = 1$. Similarly, when $p_U < \frac{1}{2}$ the θ-optimal decision is $d^*_{b,\theta}$ for all $\theta \in \Theta_0$ and so $d_{b,mmr} = d_{b,mm} = 0$. It remains to consider the case in which $p_L \leq \frac{1}{2}$ and $p_U \geq \frac{1}{2}$ both hold. It is then straightforward to deduce that

$$d_{b,mmr} = \mathbb{I}[\tfrac{1}{2} - p_L \leq p_U - \tfrac{1}{2}] = \mathbb{I}[1 \leq p_L + p_U] = d_{b,mm}\,.$$

B.3 Non-equivalence of Minimax and Minimax Regret Forecasts when $M \geq 2$

Unlike the binary case ($M = 1$), minimax and minimax regret forecasts are no longer equal for classification loss when $M \geq 2$. To see this, consider an example with $M = 3$ in which $\Theta_0 = \{\theta_1, \theta_2, \theta_3\}$ with $\theta_1 = (\frac{1}{2}, \frac{1}{2}, 0, 0)'$, $\theta_2 = (\frac{1}{3}, \frac{1}{3}, 0, \frac{1}{3})'$, and $\theta_3 = (\frac{1}{5}, \frac{1}{5}, 0, \frac{4}{5})'$, where we identify each parameter with its vector of forecast probabilities for the outcomes in the set $\mathcal{D} = \{0, 1, 2, 3\}$. The θ-optimal forecasts for classification loss are $d^*_{c,\theta_1} \in \{0, 1\}$ (i.e., both $d^*_{\theta_1,c} = 0$ and $d^*_{c,\theta_1} = 1$ are θ-optimal forecasts under θ_1), $d^*_{c,\theta_2} \in \{0, 1, 3\}$, and $d^*_{c,\theta_3} = 3$.

For the minimax decision, we have $\underline{p}_0 = \frac{1}{5}$, $\underline{p}_1 = \frac{1}{5}$, $\underline{p}_2 = 0$, and $\underline{p}_3 = 0$. Therefore, $d_{c,mm} \in \{0, 1\}$ is the minimax decision for classification loss and the minimax risk is $\mathcal{R}^*_{c,mm} = \frac{4}{5}$.

For the minimax regret decision, note that the regret from choosing $m = 0, 1, 2, 3$ under θ_1 is $(0, 0, \frac{1}{2}, \frac{1}{2})$. Similarly, under θ_2 and θ_3 the regrets are $(0, 0, \frac{1}{3}, 0)$ and $(\frac{3}{5}, \frac{3}{5}, \frac{4}{5}, 0)$. Therefore, $\Delta p_0 = \frac{3}{5}$, $\Delta p_1 = \frac{3}{5}$, $\Delta p_2 = \frac{4}{5}$, and $\Delta p_3 = \frac{1}{2}$. The minimax regret forecast is $d_{c,mmr} = 3$ and its maximum regret is $\mathcal{R}^*_{c,mmr} = \frac{1}{2}$.

Similarly, with $M = 2$ and $\theta_1 = (\frac{1}{2}, \frac{1}{2}, 0)'$, $\theta_2 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$, and $\theta_3 = (\frac{1}{5}, \frac{1}{5}, \frac{4}{5})'$, we have that the minimax forecast is $d_{c,mm} \in \{0, 1\}$ whereas the minimax regret forecast is $d_{c,mmr} = 2$.

C Proofs

C.1 Preliminaries

Our approach to establishing asymptotic efficiency follows Hirano and Porter (2009). First, we characterize the asymptotic representation of the forecast in the limit experiment. Second, we

show that these are optimal with respect to average excess maximum risk and regret in the limit experiment. Finally, we invoke a version of their Lemma 1 which allows us to approximate average excess maximum risk or regret with finite n by that in the limit experiment. The next two subsections describe preliminary results for steps 1 and 2 of this approach for binary and multinomial forecasts. The final subsection presents proofs of the main result.

To simplify notation, throughout the proofs we write $\Pi_n(P)$ for the posterior $\Pi_n(P|X_n)$, $\mathrm{d}\Pi_n$ in place of $\mathrm{d}\Pi_n(P|X_n)$. We adopt the convention that $+\infty \times 0 = 0$. We also require a limiting counterpart to excess maximum risk and regret criteria. To this end, for any sequence $\{d_n\}_{n\geq 1} \in \mathbb{D}$, $P_0 \in \mathcal{P}$, and perturbation direction $h \in \mathbb{R}^k$, we define local asymptotic excess maximum risk as

$$\mathcal{L}_{mm}(\{d_n\}_{n\geq 1}; P_0, h) = \lim_{n\to\infty} \mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}_{mm}\left(d_n, P_0 + n^{-1/2}h; X_n\right)\right].$$

Local asymptotic excess maximum regret $\mathcal{L}_{mmr}(\{d_n\}_{n\geq 1}; P_0, h)$ is defined similarly, replacing excess maximum risk $\Delta\mathcal{R}_{mm}$ in the above display with excess maximum regret $\Delta\mathcal{R}_{mmr}$. The local asymptotic excess maximum risk and regret of $\{d_n\}_{n\geq 1} \in \mathbb{D}$ will only depend on $\{d_n\}_{n\geq 1}$ through its asymptotic representation d^∞. Note the form of d^∞ may depend on P_0, but we suppress this dependence to simplify notation. We can therefore write

$$\begin{aligned}\mathcal{L}_{mm}(\{d_n\}_{n\geq 1}; P_0, h) &= \mathcal{L}^\infty_{mm}(d^\infty; P_0, h)\\ \mathcal{L}_{mmr}(\{d_n\}_{n\geq 1}; P_0, h) &= \mathcal{L}^\infty_{mmr}(d^\infty; P_0, h)\end{aligned}$$

for some functionals $\mathcal{L}^\infty_{mm}$ and $\mathcal{L}^\infty_{mmr}$. We say that d^∞_* is *optimal for average local asymptotic excess maximum risk* in the limit experiment if it is a flat prior Bayes rule:

$$\int \mathcal{L}^\infty_{mm}(d^\infty_*; P_0, h)\,\mathrm{d}h = \inf_{d^\infty} \int \mathcal{L}^\infty_{mm}(d^\infty; P_0, h)\,\mathrm{d}h$$

where the infimum on the right-hand side is taken over all such (possibly randomized) $\mathcal{D}$-valued forecasts $d^\infty(Z, U)$ in the limit experiment. Optimality for average local asymptotic excess maximum regret in the limit experiment is defined similarly. We sometimes simply say *optimal in the limit experiment* when the notion of optimality (local asymptotic minimax risk or regret) is obvious form the context.

C.2 Supplementary Lemmas: Binary Forecasts

For binary forecasts, both $\Delta\mathcal{R}_{mm}(d, P)$ and $\Delta\mathcal{R}_{mmr}(d, P)$ can be written as linear functions of d. Therefore, local asymptotic excess maximum risk and regret depend on $\{d_n\}_{n\geq 1} \in \mathbb{D}$ only through $\lim_{n\to\infty} \mathbb{E}_{P_{n,h}}[d_n(X_n)]$ which takes the form $\mathbb{E}_h[d^\infty(Z)]$ (van der Vaart, 2000, Theorem 15.1) where $\mathbb{E}_h$ denotes expectation with respect to $Z \sim N(h, I_0^{-1})$. That is not to say that the asymptotically

optimal forecast cannot be randomized. Rather, $d^\infty(z)$ represents the average (with respect to the randomization) probability that $d^\infty(Z,U)=1$ when $Z=z$.

To simplify notation, let $p_{U0} := p_U(P_0)$ and $p_{L0} := p_L(P_0)$. There are three cases to consider for the next lemma: case 1, $a_{01}p_{L0} + a_{10}p_{U0} > a_{01}$; case 2, $a_{01}p_{L0} + a_{10}p_{U0} < a_{01}$; and case 3, $a_{01}p_{L0} + a_{10}p_{U0} = a_{01}$.

Lemma C.1. *Let Assumptions 5.1 and parts (a) of Assumption 5.2 hold. Then:*

(i) $d_{b,mm}$ *has the asymptotic representation*

$$d^\infty_{b,mm}(Z) = \begin{bmatrix} 1 & \text{in case 1,} \\ 0 & \text{in case 2,} \\ \mathbb{I}\left[\mathbb{E}^*[\dot f_{P_0}[Z^* + Z]|Z] \geq 0\right] & \text{in case 3,} \end{bmatrix}$$

where $f(P) = a_{01}p_L(P) + a_{10}p_U(P)$*;*

(ii) local asymptotic excess maximum risk of $\{d_n\}_{n\geq 1} \in \mathbb{D}$ *is*

$$\mathcal{L}^\infty_{mm}(d^\infty; P_0, h) = \begin{bmatrix} +\infty \times (1 - \mathbb{E}_h[d^\infty(Z)]) & \text{in case 1,} \\ +\infty \times \mathbb{E}_h[d^\infty(Z)] & \text{in case 2,} \\ (\dot f_{P_0}[h])_+ - \mathbb{E}_h[d^\infty(Z)](\dot f_{P_0}[h]) & \text{in case 3,} \end{bmatrix}$$

where $\lim_{n\to\infty} \mathbb{E}_{P_{n,h}}[d_n(X_n)] = \mathbb{E}_h[d^\infty(Z)]$*;*

(iii) $d^\infty_{b,mm}(Z)$ *is optimal in the limit experiment.*

Proof of Lemma C.1. Part (i): As $d_{b,mm}(X_n)$ is discrete, establishing convergence in distribution of $d_{b,mm}(X_n)$ under $\{F_{n,P_{n,h}}\}_{n\geq 1}$ is equivalent to characterizing $\lim_{n\to\infty} F_{n,P_{n,h}}(d_{b,mm}(X_n) = 1)$.

For Case 1, by Assumption 5.2.1(a), for any $\varepsilon > 0$ there is a neighborhood N of P_0 upon which $|a_{01}p_L(P) + a_{10}p_U(P) - a_{01}p_{L0} - a_{10}p_{U0}| < \varepsilon$. By posterior consistency (Assumption 5.2.2(a)) and the fact that $0 \leq p_U, p_L \leq 1$, we have

$$\left| \int (a_{01}p_L(P) + a_{10}p_U(P))\, \mathrm{d}\Pi_n - a_{01}p_{L0} - a_{10}p_{U0} \right| \leq \varepsilon \Pi_n(P \in N) + 2(a_{01} + a_{10})\Pi_n(P \notin N) \xrightarrow{P_0} \varepsilon\,.$$

As ε was arbitrary, $\int (a_{01}p_L(P) + a_{10}p_U(P))\, \mathrm{d}\Pi_n \xrightarrow{P_0} a_{01}p_{L0} + a_{10}p_{U0}$. Therefore, $d_{b,mm}(X_n) \xrightarrow{P_0} 1$. As $\{F_{n,P_0}\}_{n\geq 1}$ and $\{F_{n,P_{n,h}}\}_{n\geq 1}$ are contiguous by Le Cam's first lemma and Assumption 5.1, it follows that $d_{b,mm}(X_n) \xrightarrow{P_{n,h}} 1$ for any $h \in \mathbb{R}^k$. Case 2 follows similarly. For Case 3, we may write

$$d_{b,mm}(X_n) = \mathbb{I}\left[\int a_{01}\sqrt{n}(p_L(P) - p_{L0}) + a_{10}\sqrt{n}(p_U(P) - p_{U0})\, \mathrm{d}\Pi_n \geq 0 \right].$$

By Assumption 5.2.3(a) with $A = \{(x, y) : a_{01}x + a_{10}y \geq 0\}$, we have

$$\lim_{n\to\infty} F_{n,P_{n,h}}(d_{b,mm}(X_n) = 1) = \mathbb{P}_h\left(\mathbb{E}^*[a_{01}\dot{p}_{L,P_0}[Z^* + Z] + a_{10}\dot{p}_{U,P_0}[Z^* + Z]|Z] \geq 0\right).$$

Part (ii): The excess maximum risk of $d \in \mathcal{D}$ is

$$\Delta\mathcal{R}_{mm}(d, P) = d(a_{01} - a_{01}p_L(P) - a_{10}p_U(P)) - (a_{01} - a_{01}p_L(P) - a_{10}p_U(P))_-,$$

where $a_- = \min\{a, 0\}$. For Case 1, for all n large enough we have

$$\mathbb{E}_{P_{n,h}}[\sqrt{n}\Delta\mathcal{R}_{mm}(d_n(X_n), P_{n,h})] = \sqrt{n} \times (a_{01}p_L(P_{n,h}) + a_{10}p_U(P_{n,h}) - a_{01}) \times \left(1 - \mathbb{E}_{P_{n,h}}[d_n(X_n)]\right)$$

where $\liminf_{n\to\infty}(a_{01}p_L(P_{n,h}) + a_{10}p_U(P_{n,h}) - a_{01}) > 0$ and $\mathbb{E}_{P_{n,h}}[d_n(X_n)] \to \mathbb{E}_h[d^\infty(Z)]$. Case 2 follows by similar arguments. For Case 3, rearranging slightly we have

$$\begin{aligned}\mathbb{E}_{P_{n,h}}[\sqrt{n}\Delta\mathcal{R}_{mm}(d_n(X_n), P_{n,h})] = {} & \sqrt{n}(a_{01}p_L(P_{n,h}) + a_{10}p_U(P_{n,h}) - a_{01}p_{L0} - a_{10}p_{U0})_+ \\ & - \mathbb{E}_{P_{n,h}}[d_n(X_n)] \times \sqrt{n}(a_{01}p_L(P_{n,h}) + a_{10}p_U(P_{n,h}) - a_{01}p_{L0} - a_{10}p_{U0})\end{aligned}$$

where $\mathbb{E}_{P_{n,h}}[d_n(X_n)] \to \mathbb{E}_h[d^\infty(Z)]$ and

$$\begin{aligned}\sqrt{n}(a_{01}p_L(P_{n,h}) + a_{10}p_U(P_{n,h}) - a_{01}p_{L0} - a_{10}p_{U0}) &\to a_{01}\dot{p}_{L,P_0}[h] + a_{10}\dot{p}_{U,P_0}[h], \\ \sqrt{n}(a_{01}p_L(P_{n,h}) + a_{10}p_U(P_{n,h}) - a_{01}p_{L0} - a_{10}p_{U0})_+ &\to (a_{01}\dot{p}_{L,P_0}[h] + a_{10}\dot{p}_{U,P_0}[h])_+\end{aligned}$$

by Assumption 5.2.1(a).

Part (iii): From part (ii), we see that $d^\infty(Z) = 1$ (almost everywhere) is optimal in Case 1 and $d^\infty(Z) = 0$ (almost everywhere) is optimal in Case 2. In Case 3, we have

$$\begin{aligned}&\int (a_{01}\dot{p}_{L,P_0}[h] + a_{10}\dot{p}_{U,P_0}[h])_+ - \mathbb{E}_h[d^\infty(Z)] \times (a_{01}\dot{p}_{L,P_0}[h] + a_{10}\dot{p}_{U,P_0}[h])\,\mathrm{d}h \\ &\propto \int\int ((a_{01}\dot{p}_{L,P_0}[h] + a_{01}\dot{p}_{U,P_0}[h])_+ - d^\infty(z) \times (a_{01}\dot{p}_{L,P_0}[h] + a_{10}\dot{p}_{U,P_0}[h]))\, e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}z\,\mathrm{d}h.\end{aligned}$$

Swapping the order of integration and minimizing pointwise in z, we obtain

$$d^\infty(z) = \mathbb{I}\left[\int (a_{10}\dot{p}_{L,P_0}[h] + a_{01}\dot{p}_{U,P_0}[h])e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h \geq 0\right].$$

Equivalently, $d^\infty(z) = \mathbb{I}[\mathbb{E}^*[a_{10}\dot{p}_{L,P_0}[Z^* + Z] + a_{01}\dot{p}_{U,P_0}[Z^* + Z]|Z = z] \geq 0]$. This is the same asymptotic representation as was derived in Part (i). □

Let $a = \frac{a_{01}}{a_{01}+a_{10}}$ There are four cases to consider for the next lemma, namely: case 1, $p_{L0}+p_{U0} >$

$2a$; case 2, $p_{L0} + p_{U0} < 2a$; case 3, $p_{L0} + p_{U0} = 2a$ and $p_{U0} > a$; and case 4, $p_{L0} = p_{U0} = a$.

Lemma C.2. *Let Assumption 5.1 and 5.2 hold. Then:*

(i) $d_{b,mmr}$ *has the asymptotic representation*

$$d^{\infty}_{b,mmr}(Z) = \begin{bmatrix} 1 & \text{in case 1,} \\ 0 & \text{in case 2,} \\ \mathbb{I}\left[\mathbb{E}^*[\dot{p}_{L,P_0}[Z^* + Z] + \dot{p}_{U,P_0}[Z^* + Z]|Z] \geq 0\right] & \text{in case 3,} \\ \mathbb{I}\left[\mathbb{E}^*[(\dot{p}_{L,P_0}[Z^* + Z])_- + (\dot{p}_{U,P_0}[Z^* + Z])_+|Z] \geq 0\right] & \text{in case 4;} \end{bmatrix}$$

(ii) local asymptotic excess maximum regret of $\{d_n\}_{n\geq 1} \in \mathbb{D}$ *is*

$$\begin{aligned} &\mathcal{L}^{\infty}_{mmr}(d^{\infty}; P_0, h) \\ &= \begin{bmatrix} +\infty \times (1 - \mathbb{E}_h[d^{\infty}(Z)]) & \text{in case 1,} \\ +\infty \times \mathbb{E}_h[d^{\infty}(Z)] & \text{in case 2,} \\ (a_{01} + a_{10})\left((\dot{p}_{L,P_0}[h] + \dot{p}_{U,P_0}[h])_+ - \mathbb{E}_h[d^{\infty}(Z)](\dot{p}_{L,P_0}[h] + \dot{p}_{U,P_0}[h])\right) & \text{in case 3,} \\ (a_{01} + a_{10})\left(((\dot{p}_{L,P_0}[h])_- + (\dot{p}_{U,P_0}[h])_+)_+ - \mathbb{E}_h[d^{\infty}(Z)]((\dot{p}_{L,P_0}[h])_- + (\dot{p}_{U,P_0}[h])_+)\right) & \text{in case 4,} \end{bmatrix} \end{aligned}$$

where $\lim_{n\to\infty} \mathbb{E}_{P_{n,h}}[d_n(X_n)] = \mathbb{E}_h[d^{\infty}(Z)]$*;*

(iii) $d^{\infty}_{b,mmr}(Z)$ *is optimal in the limit experiment.*

Proof of Lemma C.2. Part (i): Cases 1 and 2 follow by similar arguments to the Proof of Lemma C.1(i). For Case 3, let $\kappa := p_{U0} - a = a - p_{L0}$ and note $\kappa > 0$. We have

$$\begin{aligned} F_{n,P_{n,h}}(d_{b,mmr}(X_n) = 1) &= F_{n,P_{n,h}}\left(\textstyle\int (a - p_L(P))_+ \mathrm{d}\Pi_n \leq \int (p_U(P) - a)_+ \mathrm{d}\Pi_n\right) \\ &\geq F_{n,P_{n,h}}\left(\textstyle\int (a - p_L(P))_+ \mathrm{d}\Pi_n \leq \int (p_U(P) - a)\, \mathrm{d}\Pi_n\right) \\ &= F_{n,P_{n,h}}\left(\textstyle\int (\kappa - (p_L(P) - p_{L0}))_+ \mathrm{d}\Pi_n \leq \int (p_U(P) - p_{U0})\, \mathrm{d}\Pi_n + \kappa\right). \end{aligned}$$

As $(x - y)_+ - x = \max(-y, -x)$ and hence $(x - y)_+ + y - x = \max(0, y - x)$, we can rewrite the preceding inequality as

$$\begin{aligned} &F_{n,P_{n,h}}(d_{b,mmr}(X_n) = 1) \\ &\geq F_{n,P_{n,h}}\left(\textstyle\int ((p_L(P) - p_{L0}) - \kappa)_+ \mathrm{d}\Pi_n \leq \int (p_L(P) + p_U(P) - p_{L0} - p_{U0})\, \mathrm{d}\Pi_n\right). \end{aligned} \tag{A.9}$$

By continuity of $p_L(P)$ at P_0 (by Assumption 5.2.1(a)) and posterior consistency, we can choose a neighborhood N_κ of P_0 upon which $|p_L(P) - p_L(P_0)| < \kappa$. By Assumption 5.2.2(b), there exists $\gamma > \frac{1}{2}$ such that $n^\gamma \Pi_n(P \notin N_\kappa) \xrightarrow{P_0} 0$. As $0 \leq p_L \leq 1$, we therefore have the bound

$$n^\gamma \int ((p_L(P) - p_L(P_0)) - \kappa)_+ \mathrm{d}\Pi_n \leq 2n^\gamma \Pi_n(P \notin N_\kappa) \xrightarrow{P_0} 0\,.$$

By contiguity, convergence holds under $P_{n,h}$ for all $h \in \mathbb{R}_k$. We therefore have that

$$F_{n,P_{n,h}}\left(\int (\sqrt{n}(p_L(P) - p_L(P_0)) - \sqrt{n}\kappa)_+ \mathrm{d}\Pi_n \leq n^{\gamma-\frac{1}{2}}\right) \to 1$$

for all $h \in \mathbb{R}^k$. We may therefore rewrite (A.9) as

$$\begin{aligned} F_{n,P_{n,h}}(d_{b,mmr}(X_n) = 1) &\geq F_{n,P_{n,h}}\left(n^{\gamma-\frac{1}{2}} \leq \int (p_L(P) + p_U(P) - p_{L0} - p_{U0})\,\mathrm{d}\Pi_n\right) - o(1) \\ &\geq F_{n,P_{n,h}}\left(\varepsilon \leq \int (p_L(P) + p_U(P) - p_{L0} - p_{U0})\,\mathrm{d}\Pi_n\right) - o(1) \\ &\to \mathbb{P}_h\left(\mathbb{E}^*[\dot{p}_{L,P_0}[Z^* + Z] + \dot{p}_{U,P_0}[Z^* + Z]|Z] \geq \varepsilon\right) \end{aligned}$$

for any $\varepsilon > 0$, where the final line is by Assumption 5.2.3(a) with $A = \{(x, y) : x+y \geq \varepsilon\}$. Similarly,

$$\begin{aligned} F_{n,P_{n,h}}(d_{b,mmr}(X_n) = 1) &\leq F_{n,P_{n,h}}\left(-\varepsilon \leq \int (p_L(P) + p_U(P) - p_{L0} - p_{U0})\,\mathrm{d}\Pi_n\right) + o(1) \\ &\to \mathbb{P}_h\left(\mathbb{E}^*[\dot{p}_{L,P_0}[Z^* + Z] + \dot{p}_{U,P_0}[Z^* + Z]|Z] \geq -\varepsilon\right) \end{aligned}$$

for every $\varepsilon > 0$. The desired convergence now follows by Assumption 5.2.1(b).

Finally, consider Case 4 ($p_{U0} = p_{L0} = a$). We may write

$$d_{b,mmr}(X_n) = \mathbb{I}\left[0 \leq \int (p_L(P) - p_{L0})_- + (p_U(P) - p_{U0})_+ \mathrm{d}\Pi_n\right].$$

It follows by Assumption 5.2.3(b) taking $A = \{(x, y) : x + y \geq 0\}$ that

$$\lim_{n\to\infty} F_{n,P_{n,h}}(d_{b,mmr}(X_n) = 1) \to \mathbb{P}_h\left(\mathbb{E}^*[(\dot{p}_{L,P_0}[Z^* + Z])_- + (\dot{p}_{U,P_0}[Z^* + Z])_+|Z] \geq 0\right).$$

Part (ii): First note that excess maximum regret of $d \in \mathcal{D}$ is

$$\begin{aligned} \Delta\mathcal{R}_{mmr}(d, P) = {} & d(a_{01} - (a_{01} + a_{10})p_L(P))_+ + (1 - d)((a_{01} + a_{10})p_U(P) - a_{01})_+ \\ & - (a_{01} - (a_{01} + a_{10})p_L(P))_+ \wedge ((a_{01} + a_{10})p_U(P) - a_{01})_+ . \end{aligned}$$

For Case 1, note $((a_{01}+a_{10})p_U(P_{n,h})-a_{01})_+ = (a_{01}+a_{10})p_U(P_{n,h})-a_{01} > 0$ holds for n sufficiently large because $p_{U0} > a$ in this case. Moreover, in this case $a_{01}-(a_{01}+a_{10})p_{L0} < (a_{01}+a_{10})p_{U0}-a_{01}$, so the term $\sqrt{n}((a_{01}+a_{10})p_U(P_{n,h})-a_{01})_+$ will dominate the term $\sqrt{n}(a_{01}-(a_{01}+a_{10})p_L(P_{n,h}))_+$ asymptotically. It follows that for any $\{d_n\}_{n\geq 1} \in \mathbb{D}$,

$$\lim_{n\to\infty} \mathbb{E}_{P_{n,h}}[\sqrt{n}\Delta\mathcal{R}_{mmr}(d_n(X_n), P_{n,h})] = \lim_{n\to\infty} \sqrt{n} \times (p_U(P_{n,h}) - a) \times \left(1 - \mathbb{E}_{P_{n,h}}[d_n(X_n)]\right)$$

where $p_U(P_{n,h}) \to p_{U0} > a$ and $\mathbb{E}_{P_{n,h}}[d_n(X_n)] \to \mathbb{E}_h[d^\infty(Z)]$. Case 2 follows similarly.

For Case 3, first note that for n sufficiently large we have

$$(a_{01} - (a_{01} + a_{10})p_L(P_{n,h}))_+ = a_{01} - (a_{01} + a_{10})p_L(P_{n,h})\,,$$
$$((a_{01} + a_{10})p_U(P_{n,h}) - a_{01})_+ = (a_{01} + a_{10})p_U(P_{n,h}) - a_{01}\,.$$

Letting $\sqrt{n}(a_{01} - (a_{01} + a_{10})p_{L0}) = \sqrt{n}((a_{01} + a_{10})p_{U0} - a_{01}) = \sqrt{n}\kappa$ where $\kappa > 0$ and taking n sufficiently large, we therefore obtain

$$\begin{aligned}
&\mathbb{E}_{P_{n,h}}[\sqrt{n}\Delta\mathcal{R}_{mmr}(d_n(X_n), P_{n,h})] \\
&= \sqrt{n} \times \mathbb{E}_{P_{n,h}}[d_n(X_n)]\,(\kappa - (a_{01} + a_{10})(p_L(P_{n,h}) - p_{L0})) \\
&\quad + \sqrt{n} \times (1 - \mathbb{E}_{P_{n,h}}[d_n(X_n)])\,(\kappa + (a_{01} + a_{10})(p_U(P_{n,h}) - p_{U0})) \\
&\quad - \sqrt{n} \times ((\kappa - (a_{01} + a_{10})(p_L(P_{n,h}) - p_{L0})) \wedge (\kappa + (a_{01} + a_{10})(p_U(P_{n,h}) - p_{U0}))) \\
&= (a_{01} + a_{10})\Big(\mathbb{E}_{P_{n,h}}[d_n(X_n)] \times -\sqrt{n}(p_L(P_{n,h}) - p_{L0}) + (1 - \mathbb{E}_{P_{n,h}}[d_n(X_n)]) \times \sqrt{n}(p_U(P_{n,h}) - p_{U0}) \\
&\quad - \big((-\sqrt{n}(p_L(P_{n,h}) - p_{L0})) \wedge (\sqrt{n}(p_U(P_{n,h}) - p_{U0}))\big)\Big)\,,
\end{aligned}$$

which converges to

$$(a_{01} + a_{10})\Big(-\mathbb{E}_h[d^\infty(Z)]\dot{p}_{L,P_0}[h] + (1 - \mathbb{E}_h[d^\infty(Z)])\dot{p}_{U,P_0}[h] - ((-\dot{p}_{L,P_0}[h]) \wedge (\dot{p}_{U,P_0}[h]))\Big)\,.$$

by Assumption 5.2.1(a). The stated form now follows because $x - ((-y) \wedge x) = (x + y)_+$.

Finally, for Case 4 $a_{01} - (a_{01} + a_{10})p_{L0} = (a_{01} + a_{10})p_{U0} - a_{01} = 0$. By similar logic to Case 3.,

$$\begin{aligned}
&\mathbb{E}_{P_{n,h}}[\sqrt{n}\Delta\mathcal{R}_{mmr}(d_n(X_n), P_{n,h})] \\
&= (a_{01} + a_{10}) \times \sqrt{n} \times \Big(\mathbb{E}_{P_{n,h}}[d_n(X_n)]\,(-(p_L(P_{n,h}) - p_{L0}))_+ \\
&\quad + (1 - \mathbb{E}_{P_{n,h}}[d_n(X_n)])\,(p_U(P_{n,h}) - p_{U0})_+ - (-(p_L(P_{n,h}) - p_{L0}))_+ \wedge (p_U(P_{n,h}) - p_{U0})_+\Big)\,,
\end{aligned}$$

which converges to

$$(a_{01} + a_{10})\Big(\mathbb{E}_h[d^\infty(Z)](-\dot{p}_{L,P_0}[h])_+ + (1 - \mathbb{E}_h[d^\infty(Z)])(\dot{p}_{U,P_0}[h])_+ - ((-\dot{p}_{L,P_0}[h])_+ \wedge (\dot{p}_{U,P_0}[h])_+)\Big)\,.$$

again by Assumption 5.2.1(a). The result follows from $a - (b \wedge a) = (a - b)_+$ and $-(-a)_+ = a_-$.

Part (iii): From part (ii), we see that $d^\infty(Z) = 1$ (almost everywhere) is optimal in Case 1 and $d^\infty(Z) = 0$ (almost everywhere) is optimal in Case 2. In Case 3, by similar arguments to the proof of Lemma C.2(iii) we see that average asymptotic excess maximum regret is minimized with

$$d^\infty_{P_0}(z) = \mathbb{I}\left[\int (\dot{p}_{L,P_0}[h] + \dot{p}_{U,P_0}[h])e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h \geq 0\right]\,,$$

whereas a minimizing choice in Case 4 is

$$d^{\infty}_{P_0}(z) = \mathbb{I}\left[\int ((\dot{p}_{L,P_0}[h])_- + (\dot{p}_{U,P_0}[h])_+)e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h \geq 0\right].$$

□

C.3 Supplementary Lemmas: Multinomial Forecasts

For multinomial forecasts, the excess maximum risk $\Delta\mathcal{R}_{mm}(d,P)$ and regret $\Delta\mathcal{R}_{mmr}(d,P)$ are linear in the indicator functions $\mathbb{I}[d=m]$ for $m=0,\ldots,M$. Therefore, local asymptotic excess maximum risk and regret depend on $\{d_n\}_{n\geq 1}\in\mathbb{D}$ through $\lim_{n\to\infty}\mathbb{E}_{P_{n,h}}[\mathbb{I}(d_n(X_n)=m)]$ which can be written $\mathbb{E}_h[d^{\infty}_m(Z)]$ for each m (van der Vaart, 2000, Theorem 15.1). The term $d^{\infty}_m(z)$ represents the average (with respect to randomization) probability that $d^{\infty}(Z,U)=m$ when $Z=z$.

Deriving the asymptotic representation of $d_{c,mm}(X_n)$ requires a tie-breaking rule, so in the derivation below we take the smallest element of the set of maximizers. To simplify notation, let $\underline{p}_{m0} := \underline{p}_m(P_0)$. It is without loss of generality to reorder the indices so that $\underline{p}_{00}\geq\underline{p}_{10}\geq\ldots\geq\underline{p}_{M0}$. There are two cases, namely: case 1, $\underline{p}_{00}>\underline{p}_{10}$; and case 2, $\underline{p}_{00}=\underline{p}_{10}=\ldots=\underline{p}_{k0}$ for some $k\in\{1,\ldots,M\}$ with $\underline{p}_{k0}>\underline{p}_{(k+1)0}$ if $k<M$.

Lemma C.3. *Let Assumptions 5.1, 5.10.1(a), 5.10.2, and 5.10.3(a) hold. Then:*

(i) $d_{c,mm}$ *has the asymptotic representation*

$$d^{\infty}_{c,mm,m}(Z) = \begin{bmatrix} 1 \text{ if } m=0 \text{ and } 0 \text{ if } m\in\{1,\ldots,M\} & \text{in case 1,}\\ \mathbb{I}[(\mathbb{E}^*[\dot{\underline{p}}_{m,P_0}[Z^*+Z]|Z] > \max_{0\leq m'\leq m-1}\mathbb{E}^*[\dot{\underline{p}}_{m',P_0}[Z^*+Z]|Z]) \text{ and} & \\ \quad (\mathbb{E}^*[\dot{\underline{p}}_{m,P_0}[Z^*+Z]|Z] \geq \max_{m+1\leq m'\leq k}\mathbb{E}^*[\dot{\underline{p}}_{m',P_0}[Z^*+Z]|Z])] & \\ \text{if } m\in\{0,\ldots,k\} \text{ and } 0 \text{ if } m\in\{k+1,\ldots,M\} & \text{in case 2,}\end{bmatrix}$$

where the maximum over an empty index is $-\infty$*;*

(ii) local asymptotic excess maximum risk of $\{d_n\}_{n\geq 1}\in\mathbb{D}$ *is*

$$\mathcal{L}^{\infty}_{mm}(d^{\infty};P_0,h) = \begin{bmatrix} +\infty\times(1-\mathbb{E}_h[d^{\infty}_0(Z)]) & \text{in case 1,}\\ \sum_{m=0}^{k}\mathbb{E}_h[d^{\infty}_m(Z)](\max_{0\leq m'\leq k}\dot{\underline{p}}_{m',P_0}[h]-\dot{\underline{p}}_{m,P_0}[h]) & \\ \quad +\infty\times(1-\sum_{m=0}^{k}\mathbb{E}_h[d^{\infty}_m(Z)]) & \text{in case 2,}\end{bmatrix}$$

where $\lim_{n\to\infty}\mathbb{E}_{P_{n,h}}[\mathbb{I}[d_n(X_n)=m]] = \mathbb{E}_h[d^{\infty}_m(Z)]$*;*

(iii) $(d^{\infty}_{c,mm,m}(Z))_{m=0}^{M}$ *is optimal in the limit experiment.*

Proof of Lemma C.3. Part (i): Case 1 follows by similar arguments to the proof of Lemma C.1. For Case 2, if $k<M$ we can deduce by continuity of the $\underline{p}_m$ and posterior consistency that

$F_{n,P_{n,h}}(d_{c,mm}(X_n) > k) \to 0$ and $F_{n,P_{n,h}}(\min_{0\le m\le k}\int \underline{p}_m(P)\,\mathrm{d}\Pi_n > \max_{m>k}\int \underline{p}_m(P)\,\mathrm{d}\Pi_n) \to 1$. Let $\min_{0\le m\le k}\int \underline{p}_m(P)\,\mathrm{d}\Pi_n > \max_{m>k}\int \underline{p}_m(P)\,\mathrm{d}\Pi_n$. For $m \in \{0, 1\ldots, k\}$, under the above tie-breaking rule we then have

$$\mathbb{I}[d_{c,mm}(X_n) = m] = \mathbb{I}\left[\int \underline{p}_m(P)\,\mathrm{d}\Pi_n > \max_{0\le m'\le m-1}\int \underline{p}_{m'}(P)\,\mathrm{d}\Pi_n\right] \times \mathbb{I}\left[\int \underline{p}_m(P)\,\mathrm{d}\Pi_n \ge \max_{m+1\le m'\le k}\int \underline{p}_{m'}(P)\,\mathrm{d}\Pi_n\right].$$

As $\underline{p}_{00} = \underline{p}_{10} = \ldots = \underline{p}_{k0}$, we may rewrite the previous expression as

$$\mathbb{I}[d_{c,mm}(X_n) = m] = \mathbb{I}\left[\int \sqrt{n}(\underline{p}_m(P) - \underline{p}_{m0})\,\mathrm{d}\Pi_n > \max_{0\le m'\le m-1}\int \sqrt{n}(\underline{p}_{m'}(P) - \underline{p}_{m'0})\,\mathrm{d}\Pi_n\right] \times \mathbb{I}\left[\int \sqrt{n}(\underline{p}_m(P) - \underline{p}_{m0})\,\mathrm{d}\Pi_n \ge \max_{m+1\le m'\le k}\int \sqrt{n}(\underline{p}_{m'}(P) - \underline{p}_{m'0})\,\mathrm{d}\Pi_n\right].$$

Therefore by Assumptions 5.10.1(a) and 5.10.3(a) with $A = \{(x_0, x_1, \ldots, x_M) : x_m > x_{m'}$ if $m' \in \{0, \ldots, m-1\}$ and $x_m \ge x_{m'}$ if $m' \in \{m+1, \ldots, k\}\}$, we have

$$\lim_{n\to\infty} F_{n,P_{n,h}}(d_{c,mm}(X_n) = m) = \mathbb{P}_h\Big(\Big(\mathbb{E}^*[\dot{\underline{p}}_m[Z^* + Z]|Z] > \max_{0\le m'\le m-1}\mathbb{E}^*[\dot{\underline{p}}_{m'}[Z^* + Z]|Z]\Big) \text{ and } \Big(\mathbb{E}^*[\dot{\underline{p}}_m[Z^* + Z]|Z] \ge \max_{m+1\le m'\le k}\mathbb{E}^*[\dot{\underline{p}}_{m'}[Z^* + Z]|Z]\Big)\Big).$$

Part (ii): The excess maximum risk of $d \in \mathcal{D}$ is

$$\Delta\mathcal{R}_{mm}(d, P) = \sum_{m=0}^{M}\mathbb{I}[d = m]\left(\max_{0\le m'\le M}\underline{p}_{m'}(P) - \underline{p}_m(P)\right).$$

For Case 1, by continuity of $\underline{p}_m(\cdot)$ for all m (under Assumption 5.10.1(a)) we have $\underline{p}_m(P_{n,h}) \to \underline{p}_{m0}$ for all m and $\max_{0\le m'\le M}\underline{p}_{m'}(P_{n,h}) \to \underline{p}_{00}$. Then for all n sufficiently large,

$$\mathbb{E}_{P_{n,h}}[\sqrt{n}\Delta\mathcal{R}_{mm}(d_n(X_n), P_{n,h})] = \sqrt{n}\sum_{m=1}^{M}\mathbb{E}_{P_{n,h}}\left[\mathbb{I}[d_n(X_n) = m]\right]\left(\underline{p}_0(P_{n,h}) - \underline{p}_m(P_{n,h})\right)$$

where $\liminf_{n\to\infty}(\underline{p}_0(P_{n,h}) - \underline{p}_m(P_{n,h})) > 0$ for $m \ge 1$ and $\mathbb{E}_{P_{n,h}}[\mathbb{I}[d_n(X_n) = m]] \to \mathbb{E}_h[d_m^\infty(Z)]$. Now consider Case 2. Again by continuity, for n sufficiently large we have

$$\mathbb{E}_{P_{n,h}}[\sqrt{n}\Delta\mathcal{R}_{mm}(d_n(X_n), P_{n,h})] = \sqrt{n}\sum_{m=1}^{k}\mathbb{E}_{P_{n,h}}\left[\mathbb{I}[d_n(X_n) = m]\right]\left(\max_{0\le m'\le k}\underline{p}_{m'}(P_{n,h}) - \underline{p}_m(P_{n,h})\right) + \sqrt{n}\sum_{m=k+1}^{M}\mathbb{E}_{P_{n,h}}\left[\mathbb{I}[d_n(X_n) = m]\right]\left(\max_{0\le m'\le k}\underline{p}_{m'}(P_{n,h}) - \underline{p}_m(P_{n,h})\right)$$

where the second sum is zero if $k = M$. If $k < M$, by similar arguments to Case 1 we have

$$\sqrt{n} \sum_{m=k+1}^{M} \mathbb{E}_{P_{n,h}} [\mathbb{I}[d_n(X_n) = m]] \left(\max_{0 \le m' \le k} \underline{p}_{m'}(P_{n,h}) - \underline{p}_m(P_{n,h}) \right) \to +\infty \times \sum_{m=k+1}^{M} \mathbb{E}_h[d_m^\infty(Z)] \,.$$

Moreover, for $m \le k$ by Assumption 5.10.1(a) we have

$$\begin{aligned} \sqrt{n} \left(\max_{0 \le m' \le k} \underline{p}_{m'}(P_{n,h}) - \underline{p}_m(P_{n,h}) \right) &= \left(\max_{0 \le m' \le k} (\sqrt{n}(\underline{p}_{m'}(P_{n,h}) - \underline{p}_{m'0})) - \sqrt{n}(\underline{p}_m(P_{n,h}) - \underline{p}_{m0}) \right) \\ &\to \max_{0 \le m' \le k} \dot{\underline{p}}_{m',P_0}[h] - \dot{\underline{p}}_{m,P_0}[h] \,. \end{aligned}$$

Part (iii): For Case 1, from part (ii), we see that $d_0^\infty = 1$ (almost everywhere) is optimal. For Case 2, from part (ii), we see that $d_m^\infty = 0$ (almost everywhere) is optimal for all $m > k$. For the remaining values of m, we have

$$\begin{aligned} &\int \sum_{m=0}^{k} \mathbb{E}_h[d_m^\infty(Z)] \left(\max_{0 \le m' \le k} \dot{\underline{p}}_{m',P_0}[h] - \dot{\underline{p}}_{m,P_0}[h] \right) \mathrm{d}h \\ &\propto \int \int d_{m,P_0}^\infty(z) \left(\max_{0 \le m' \le k} \dot{\underline{p}}_{m',P_0}[h] - \dot{\underline{p}}_{m,P_0}[h] \right) e^{-\frac{1}{2}(z-h)' I_0 (z-h)} \, \mathrm{d}z \, \mathrm{d}h \,. \end{aligned}$$

Changing the order of integration and minimizing pointwise in z, we see that if $M(z)$ denotes the set of maximizers of

$$\int \dot{\underline{p}}_{m,P_0}[h] e^{-\frac{1}{2}(z-h)' I_0 (z-h)} \, \mathrm{d}h \,,$$

then setting $d_m^\infty(z) = 0$ for $m \notin M(z)$ and $d_m^\infty(z) \ge 0$ for $m \in M(z)$ with $\sum_{m \in M(z)} d_m^\infty(z) = 1$ is optimal. The tie-breaking rule used in part (i) is a special case with $d_m^\infty(z) = 1$ if $m = \min M(z)$. □

Characterizing $d_{c,mmr}(X_n)$ again requires a tie-breaking rule. In the derivation below we take the smallest element of the set of minimizers. To simplify notation, let $\tau_m(P) = \Delta p_m(P)$ and $\tau_{m0} = \tau_m(P_0)$ for $m = 0, 1, \ldots, M$. Without loss of generality, reorder the indices so that $\tau_{00} \le \tau_{10} \le \ldots \le \tau_{M0}$. There are two cases, namely: case 1, $\tau_{00} < \tau_{10}$; and case 2, $\tau_{00} = \tau_{10} = \ldots = \tau_{k0}$ for some $k \in \{1, \ldots, M\}$ with $\tau_{k0} < \tau_{(k+1)0}$ if $k < M$.

Lemma C.4. *Let Assumptions 5.1, 5.10.1(b), 5.10.2, and 5.10.3(b) hold. Then:*

(i) $d_{c,mmr}$ has the asymptotic representation

$$d_{c,mmr,m}^\infty(Z) = \begin{cases} 1 \text{ if } m = 0 \text{ and } 0 \text{ if } m \in \{1, \ldots, M\} & \text{in case 1,} \\ \mathbb{P}_h((\mathbb{E}^*[\dot{\tau}_m[Z^* + Z]|Z] < \min_{0 \le m' \le m-1} \mathbb{E}^*[\dot{\tau}_{m'}[Z^* + Z]|Z]) \text{ and} \\ \quad (\mathbb{E}^*[\dot{\tau}_m[Z^* + Z]|Z] \le \min_{m+1 \le m' \le k} \mathbb{E}^*[\dot{\tau}_{m'}[Z^* + Z]|Z])) \\ \text{if } m \in \{0, \ldots, k\} \text{ and } 0 \text{ if } m \in \{k+1, \ldots, M\} & \text{in case 2,} \end{cases}$$

where the minimum over an empty index is $+\infty$;

(ii) local asymptotic excess maximum risk of $\{d_n\}_{n\geq 1} \in \mathbb{D}$ is

$$\mathcal{L}^{\infty}_{mmr}(d^{\infty}; P_0, h) = \begin{bmatrix} +\infty \times (1 - \mathbb{E}_h[d_0^{\infty}(Z)]) & \text{in case 1,} \\ \sum_{m=0}^{k} \mathbb{E}_h[d_m^{\infty}(Z)](\dot{\tau}_{m,P_0}[h] - \min_{0\leq m'\leq k} \dot{\tau}_{m',P_0}[h]) & \\ \quad +\infty \times (1 - \sum_{m=0}^{k} \mathbb{E}_h[d_m^{\infty}(Z)]) & \text{in case 2,} \end{bmatrix}$$

where $\lim_{n\to\infty} \mathbb{E}_{P_{n,h}}[\mathbb{I}(d_n(X_n) = m)] = \mathbb{E}_h[d_m^{\infty}(Z)]$;

(iii) $(d^{\infty}_{c,mmr,m}(Z))_{m=0}^{M}$ is optimal in the limit experiment.

Proof of Lemma C.4. Follows by similar arguments to the proof of Lemma C.3. □

C.4 Main results

Theorems 5.4 and 5.11 are proved using the following lemma, which is a very slight generalization of Lemma 1 of Hirano and Porter (2009). We include a proof for completeness. It applies to both minimax risk and regret criteria, so we drop the subscripts *mm* and *mmr* on $\mathcal{L}$, $\Delta\mathcal{B}$, and $\mathcal{R}$.

Lemma C.5. *Let $\mathcal{L}(\{d_n\}_{n\geq 1}; P_0, h) = \mathcal{L}^{\infty}(d^{\infty}; P_0, h)$ hold for every $P_0 \in \mathcal{P}$, $h \in \mathbb{R}^k$, and $\{d_n\}_{n\geq 1} \in \mathbb{D}$, where d^{∞} denotes the asymptotic representation of $\{d_n\}_{n\geq 1} \in \mathbb{D}$, and let the prior Π have a strictly positive, continuously differentiable density π on $\mathcal{P}$. Then: (i) for any $\{d_n\}_{n\geq 1} \in \mathbb{D}$,*

$$\liminf_{n\to\infty} \Delta\mathcal{B}^n(d_n; P_0, \pi) \geq \pi(P_0) \inf_{d^{\infty}} \int \mathcal{L}^{\infty}(d^{\infty}; P_0, h)\, \mathrm{d}h$$

(ii) If, in addition, $\{d_n^*\}_{n\geq 1} \in \mathbb{D}$ and its asymptotic representation d_*^{∞} solves

$$\int \mathcal{L}^{\infty}(d_*^{\infty}; P_0, h)\, \mathrm{d}h = \inf_{d^{\infty}} \int \mathcal{L}^{\infty}(d^{\infty}; P_0, h)\, \mathrm{d}h\,,$$

and d_n^* satisfies

$$\limsup_{n\to\infty} \Delta\mathcal{B}^n(d_n^*; P_0, \pi) \leq \int \limsup_{n\to\infty} \mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}\left(d_n^*, P_{n,h}; X_n\right)\right] \pi\left(P_{n,h}\right)\, \mathrm{d}h\,, \tag{A.10}$$

then:

$$\lim_{n\to\infty} \Delta\mathcal{B}^n(d_n^*; P_0, \pi) = \pi(P_0) \inf_{d^{\infty}} \int \mathcal{L}^{\infty}(d^{\infty}; P_0, h)\, \mathrm{d}h\,,$$

and hence

$$\lim_{n\to\infty} \Delta\mathcal{B}^n(d_n^*; P_0, \pi) = \inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n(d_n; P_0, \pi)\,.$$

Remark C.6. *By the reverse Fatou lemma, condition (A.10) holds if there exists a non-negative function $g(h)$ with $\mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}\left(d_n^*, P_{n,h}; X_n\right)\right] \pi\left(P_{n,h}\right) \leq g(h)$ for each n and $\int g(h)\, \mathrm{d}h < \infty$.*

Proof of Lemma C.5. Part (i): follows by Fatou's lemma and definition of $\Delta\mathcal{B}^n(d_n; P_0, \pi)$:

$$\begin{aligned}\liminf_{n\to\infty} \Delta\mathcal{B}^n(d_n; P_0, \pi) &\geq \int \liminf_{n\to\infty} \mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}\left(d_n, P_0 + n^{-1/2}h; X_n\right)\right] \pi\left(P_0 + n^{-1/2}h\right) \mathrm{d}h \\ &= \pi(P_0) \int \mathcal{L}^\infty(d^\infty; P_0, h)\, \mathrm{d}h\,,\end{aligned}$$

where d^∞ denotes the asymptotic representation of $\{d_n\}_{n\geq 1} \in \mathbb{D}$.

Part (ii): By condition (A.10) and optimality of d_*^∞ in the limit experiment, we have

$$\begin{aligned}\limsup_{n\to\infty} \Delta\mathcal{B}^n(d_n^*; P_0, \pi) &\leq \int \limsup_{n\to\infty} \mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}\left(d_n^*, P_{n,h}; X_n\right)\right] \pi\left(P_{n,h}\right) \mathrm{d}h \\ &= \pi(P_0) \int \mathcal{L}^\infty(d_*^\infty; P_0, h)\, \mathrm{d}h \\ &= \pi(P_0) \inf_{d^\infty} \int \mathcal{L}^\infty(d^\infty; P_0, h)\, \mathrm{d}h\,.\end{aligned}$$

Combining with part (i) applied to $\{d_n^*\}_{n\geq 1}$, we obtain

$$\lim_{n\to\infty} \Delta\mathcal{B}^n(d_n^*; P_0, \pi) = \pi(P_0) \inf_{d^\infty} \int \mathcal{L}^\infty(d^\infty; P_0, h)\, \mathrm{d}h\,.$$

The final result is immediate from part (i). □

Proof of Theorem 5.4. Part (i): First note that as $\tilde{d}_n$ is binary, establishing convergence in distribution under $\{F_{n,P_{n,h}}\}_{n\geq 1}$ is equivalent to characterizing $\lim_{n\to\infty} F_{n,P_{n,h}}(\tilde{d}_n(X_n) = 1)$. Lemma C.1(i) establishes that $d_{b,mm}$ converges in distribution along every sequence $\{F_{n,P_{n,h}}\}_{n\geq 1}$. Asymptotic equivalence of $\tilde{d}_n$ and $d_{b,mm}$ implies $\lim_{n\to\infty} F_{n,P_{n,h}}(\tilde{d}_n(X_n) = 1) = \lim_{n\to\infty} F_{n,P_{n,h}}(d_{b,mm}(X_n) = 1)$ for all $h \in \mathbb{R}^k$ and all $P_0 \in \mathcal{P}$. Therefore, $\tilde{d}_n$ has the same asymptotic representation as $d_{b,mm}$ from Lemma C.1(i). As this asymptotic representation is optimal in the limit experiment (cf. Lemma C.1(iii)) and d_n satisfies condition (A.10) by assumption, the desired conclusion now follows by Lemma C.5.

Part (ii): Follows similarly by Lemmas C.2 and C.5. □

Proof of Proposition 5.8. Part (i): By Fatou's lemma and definition of $\Delta\mathcal{B}^n_{b,mm}(d_n; P_0, \pi)$, we have

$$\begin{aligned}\liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(\tilde{d}_n; P_0, \pi) &\geq \int \liminf_{n\to\infty} \mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}_{mm}\left(\tilde{d}_n, P_0 + n^{-1/2}h; X_n\right)\right] \pi\left(P_0 + n^{-1/2}h\right) \mathrm{d}h \\ &= \pi(P_0) \int \mathcal{L}^\infty_{mm}(\tilde{d}^\infty; P_0, h)\, \mathrm{d}h\,,\end{aligned}$$

where $\tilde{d}^\infty$ denotes the asymptotic representation of $\{\tilde{d}_n\}_{n\geq 1} \in \mathbb{D}$ and $\pi(P_0) > 0$. By the proof of

Theorem 5.4 we also have

$$\inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(d_n; P_0, \pi) = \lim_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(d_{b,mm}; P_0, \pi) = \pi(P_0) \int \mathcal{L}^\infty_{mm}(d^\infty_{b,mm}; P_0, h)\,\mathrm{d}h\,.$$

Therefore, it suffices to show that

$$\int \mathcal{L}^\infty_{mm}(\tilde{d}^\infty; P_0, h)\,\mathrm{d}h > \int \mathcal{L}^\infty_{mm}(d^\infty_{b,mm}; P_0, h)\,\mathrm{d}h\,. \tag{A.11}$$

First, suppose that $a_{01}p_L(P_0) + a_{10}p_U(P_0) \neq a_{01}$. This corresponds to Cases 1 and 2 of Lemma C.1. As asymptotic equivalence fails, we have

$$\lim_{n\to\infty} F_{n,P_{n,h_*}}(\tilde{d}_n(X_n) = 1) \neq \lim_{n\to\infty} F_{n,P_{n,h_*}}(d_{b,mm}(X_n) = 1)$$

for some $P_0 \in \mathcal{P}$ and h_* in $\mathbb{R}^k$. We may restate the above display in terms of the asymptotic representations:

$$\mathbb{E}_{h_*}[\tilde{d}^\infty(Z)] \neq \mathbb{E}_{h_*}[d^\infty_{b,mm}(Z)]\,.$$

By Hölder's inequality we may deduce that the functions $h \mapsto \mathbb{E}_h[\tilde{d}^\infty(Z)]$ and $h \mapsto \mathbb{E}_h[d^\infty_{b,mm}(Z)]$ are both continuous at h_*. Therefore, there exists a set $H \subset \mathbb{R}^k$ with positive Lebesgue measure upon which $\mathbb{E}_h[\tilde{d}^\infty(Z)] \neq \mathbb{E}_h[d^\infty_{b,mm}(Z)]$ for all $h \in H$.

If P_0 is as in Case 1 of Lemma C.1, then $\mathbb{E}_h[\tilde{d}^\infty(Z)] < 1$ for all $h \in H$. This, in turn, implies that $\mathcal{L}^\infty_{mm}(\tilde{d}^\infty; P_0, h) = +\infty$ for all $h \in H$. By contrast, $\mathcal{L}^\infty_{mm}(d^\infty_{b,mm}; P_0, h) = 0$ for all $h \in \mathbb{R}^k$. The proof when P_0 satisfies the conditions of Case 2 of Lemma C.1 follows similarly.

Now suppose that $a_{01}p_L(P_0) + a_{10}p_U(P_0) \neq a_{01}$, which corresponds to Case 3 of Lemma C.1. Let $f(P) = a_{01}p_L(P) + a_{10}p_U(P)$. By Lemma C.1(ii), to prove inequality (A.11) it suffices to show

$$\int \left(\tilde{d}^\infty(z) \int (\dot{f}_{P_0}[h]) e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h\right)\mathrm{d}z < \int \left(d^\infty_{b,mm}(z) \int (\dot{f}_{P_0}[h]) e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h\right)\mathrm{d}z\,.$$

The function $\tilde{d}^\infty_{b,mm}(z) = \mathbb{I}\left[\int (\dot{f}_{P_0}[h]) e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h \geq 0\right]$ maximizes

$$d(z) \times \int (\dot{f}_{P_0}[h]) e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h$$

over all $[0,1]$-valued functions of z, so the preceding inequality holds weakly. To establish a strict inequality, note the functions $\tilde{d}^\infty(z)$ and $d^\infty_{b,mm}(z)$ must disagree on a set of positive Lebesgue measure, say $\mathcal{Z}$. For each $z \in \mathcal{Z}$ we must have one of the following:

(i) $\int (\dot{f}_{P_0}[h]) e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h > 0$ and $d^\infty(z) < 1$
(ii) $\int (\dot{f}_{P_0}[h]) e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h < 0$ and $d^\infty(z) > 0$

(iii) $\int(\dot{f}_{P_0}[h])e^{-\frac{1}{2}(z-h)'I_0(z-h)}\,\mathrm{d}h = 0.$

However, the condition $\mathbb{E}^*[a_{01}\dot{p}_{L,P_0}[Z^*+Z]+a_{10}\dot{p}_{U,P_0}[Z^*+Z]|Z] \neq 0$ a.e. implies that case (iii) only holds on a set of zero Lebesgue measure. Therefore, for almost every $z \in \mathcal{Z}$ either case (i) or (ii) must hold, which establishes the desired inequality.

Part (ii): This follows by Lemma C.2 using similar arguments to Part (i). □

Proof of Theorem 5.11. The proof follows by similar arguments to Theorem 5.4, using Lemmas C.3 and C.5 for part (i) and Lemmas C.4 and C.5 for part (ii). □

C.5 Results on Computation

Proof of Proposition A.1. Dropping dependence of the L-vector b and $K \times L$ matrix G on the low-dimensional parameter ϕ, the primal problem is

$$\sup_{\pi\in\mathbb{R}^L} b'\pi \quad \text{subject to} \quad (G - r\otimes 1'_{1\times L})\pi = 0,\ 1_{1\times L}\pi - 1 = 0,\ \pi \geq 0\,,$$

where the final inequality holds element-wise. The Lagrangian is

$$\sup_{\pi\in\mathbb{R}^L}\inf_{\mu\in\mathbb{R}^K,\zeta\in\mathbb{R},\kappa\in\mathbb{R}^L_+} \mathcal{L}(\pi,\mu,\zeta,\kappa).$$

Here μ, ζ, and κ are the Lagrange multipliers on the three constraints and

$$\begin{aligned}\mathcal{L}(\pi,\mu,\zeta,\kappa) &= b'\pi + \mu'\left(G - r\otimes 1'_{1\times L}\right)\pi + \zeta\left(1_{1\times L}\pi - 1\right) + \kappa'\pi \\ &= \left(b + \left(G - r\otimes 1'_{1\times L}\right)'\mu + \zeta 1'_{1\times L} + \kappa\right)'\pi - \zeta\,.\end{aligned}$$

By duality, we have

$$\sup_{\pi}\inf_{\mu,\zeta,\kappa} \mathcal{L}(\pi,\mu,\zeta,\kappa) = \inf_{\mu,\zeta,\kappa}\sup_{\pi} \mathcal{L}(\pi,\mu,\zeta,\kappa).$$

For fixed μ, ζ, and κ, consider the problem

$$\sup_{\pi} \mathcal{L}(\pi,\mu,\zeta,\kappa) = \sup_{\pi} \underbrace{\left(b + \left(G - r\otimes 1'_{1\times L}\right)'\mu + \zeta 1'_{1\times L} + \kappa\right)'}_{=:b^*(\mu,\zeta,\kappa)'}\pi - \zeta\,.$$

This value can be made $+\infty$ by assigning arbitrarily large positive values to any element of π for which $b^*(\mu,\zeta,\kappa)$ has a positive entry, and an arbitrarily large negative value to any element of π for which $b^*(\mu,\zeta,\kappa)$ has a negative entry. The minimizing agent would therefore choose

$$\kappa^* = \kappa^*(\zeta,\mu) = -\left(\left(b_l + (G_l - r)'\mu + \zeta\right)\wedge 0\right)_{l\in\{1,\ldots,L\}} \in \mathbb{R}^L_+$$

so that all entries of $m^*(\mu, \zeta, \kappa^*)$ are non-negative:

$$m^*(\mu, \zeta, \kappa^*) = \left(\left(b_l + (G_l - r)'\mu + \zeta \right) \vee 0 \right)_{l \in \{1,\ldots,L\}} ,$$

and then choose

$$\zeta^* = \zeta(\mu) = - \max_{l \in \{1,\ldots,L\}} \left(b_l + (G_l - r)'\mu \right)$$

so that every entry of $b^*(\mu, \zeta^*, \kappa^*)$ is zero. Any $\zeta \leq \zeta^*$ will suffice for this purpose, but values of ζ strictly less than ζ^* will result in a higher value of the minimizing agent's objective. Combining the intermediate results, we obtain

$$\sup_{\pi} \inf_{\mu,\zeta,\kappa} \mathcal{L}(\pi, \mu, \zeta, \kappa) \quad = \quad \inf_{\mu} \left(\max_{l \in \{1,\ldots,L\}} b_l + \mu'(G_l - r) \right).$$

This min-max problem may be restated as a linear program by introducing an additional variable $t \in \mathbb{R}$ for the minimizing agent:

$$\begin{aligned} \inf_{\mu} \left(\max_{l \in \{1,\ldots,L\}} b_l + \mu'(G_l - r) \right) &= \inf_{\mu,t} \; t \quad \text{s.t.} \quad t \geq \left(b_l + \mu'(G_l - r) \right) , \; l = 1, \ldots, L \\ &= \inf_{\mu,t} \; t \quad \text{s.t.} \quad t 1_{L\times 1} \geq \left(b + (G' - (1_{L\times 1} \otimes r'))\mu \right) \\ &= \inf_{v} \; [0_{1\times K}, 1]\, v \quad \text{s.t.} \quad Av \leq -b \,, \end{aligned}$$

where $v = [\mu', t]' \in \mathbb{R}^{K+1}$ and $A = [G' - (1_{L\times 1} \otimes r'), -1_{L\times 1}]$. □

Proof of Proposition A.2. The dual representation follows from Csiszár and Matúš (2012). Large-δ behavior is established in Christensen and Connault (2019). □