Misspecification-Averse Estimation

Isaiah Andrews, Ricky Li, and Yucheng Shang*

Abstract

We study optimal estimation when the likelihood may be misspecified. Building on tools from the theory of decision-making under uncertainty, we analyze a class of axiomatically grounded optimality criteria which nests several existing misspecification-robust objectives. Within this class, we introduce the constrained multiplier criterion, which allows for flexible misspecification attitudes. We prove a local asymptotic minimax theorem for this criterion, extending a classical efficiency bound to a limit experiment which incorporates moment-constrained misspecification concerns. We characterize asymptotically optimal estimators as Bayes decision rules under a flat prior and an exponentially tilted likelihood that incorporates the moment constraints, and show that feasible plug-in analogs are asymptotically optimal.

1 Introduction

Researchers in economics are often concerned that their models are wrong, and about the consequences for estimation and inference. Multiple recent papers consider the problem of estimation and inference under misspecification with different optimality criteria, including minimax mean squared error (Bonhomme and Weidner, 2022) and minimax confidence interval length (Armstrong and Kolesár, 2021). Other work considers the related problem of optimal decision-making by economic agents with misspecification concerns (Hansen and Sargent, 2001, 2008; Cerreia-Vioglio et al., 2025). These contributions adopt different objective functions, and the choice among criteria has substantive consequences for the resulting procedures. A natural question is thus what objective a researcher *should* use when their

*This version: March 30, 2026. Andrews: MIT Department of Economics and NBER; iandrews@mit.edu. Li, MIT Department of Economics; rickyli@mit.edu. Shang, MIT Department of Economics; ycshang@mit.edu. We thank Ashesh Rambachan, Brit Sharoni, and Tomasz Strzalecki for collaboration on earlier project which derived related results, Drew Fudenberg for very helpful feedback, and Claude Opus 4.6 and GPT 5.4 Pro for outstanding research assistance.

model (and in particular, the model-implied mapping from parameters to data distributions) may be wrong, and what different choices of criteria imply.

The question of how to evaluate decisions in the presence of model uncertainty is a central concern of the microeconomic theory of choice under ambiguity. That literature provides axiomatic foundations for classes of preferences, linking axioms on preferences to functional representations of the decision-maker's objective (Gilboa and Schmeidler, 1989; Maccheroni et al., 2006; Cerreia-Vioglio et al., 2025). This paper applies these tools to the problem of choosing estimation criteria when the model may be misspecified, connecting axiomatic foundations for ambiguity-averse preferences to the design of statistical procedures.

Our first contribution is to introduce *constrained multiplier preferences* as an optimality criterion for estimation under misspecification. Under this criterion, a researcher evaluates an estimator by its worst-case expected loss over distributions that satisfy constraints encoding what is known about potential misspecification, penalized by a Kullback-Leibler divergence term that reflects the difference from the baseline model. We start from a broad class of misspecification-averse preferences axiomatized by Cerreia-Vioglio et al. (2025), which nests many different misspecification-robust criteria. We then discuss additional axioms that select specific subclasses within this family: constraint preferences (as in Bonhomme and Weidner, 2022), which are characterized by the certainty independence axiom of Gilboa and Schmeidler (1989); multiplier preferences (as in Hansen and Sargent, 2001), characterized by the sure thing principle following Strzalecki (2011); and the constrained multiplier class, which combines constraint and multiplier features and whose axiomatic characterization is new. Building on results from the generalized empirical likelihood literature, we further show that when the constraints can be expressed as moment conditions, the constrained multiplier preference has a computationally tractable dual representation. The moment constraints allow the researcher to express partial trust in the model: for instance, one may trust certain first order conditions implied by profit maximization, but not e.g. functional form assumptions on latent error terms.

Our second contribution is to derive optimal estimators for this criterion. We prove a local asymptotic minimax theorem that extends the classical lower bound of Hájek (1972) to the constrained multiplier objective: for convex loss, no sequence of estimators can achieve worst-case risk below our bound. Obtaining this bound requires deriving a novel limit experiment that incorporates both misspecification concerns and moment information. We then characterize estimators that attain this bound. They take the form of Bayes rules under an exponentially tilted likelihood that incorporates the moment constraints. Under squared

error loss, the structure simplifies: when no moment constraints are imposed, or when only the mean of a vector of sample average moments is constrained, the maximum likelihood estimator is optimal; when higher moments of the sample average moment vector are constrained, the optimal rule linearly adjusts the MLE using the sample average moments. We further show that plug-in finite-sample analogs of the optimal rules in the limit experiment are asymptotically optimal under regularity conditions.

Section 2 introduces the decision-theoretic framework and the axiomatization of constrained multiplier preferences. Section 3 states the asymptotic risk bound. Section 4 characterizes optimal estimators in the limit experiment and their feasible analogs. Appendix A provides further detail on the axiomatic foundations of our approach, while the remaining proofs appear in Appendix B.

2 Preferences Over Losses

Consider a researcher who will observe data X from a sample space $\mathcal{X}$. For an unknown parameter θ in a parameter space Θ, this researcher must choose an action a from a set of feasible actions $\mathcal{A}$. The researcher's objective is specified by a loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$, where $L(a, \theta)$ describes the loss from taking action a when the true parameter is θ.

Since the researcher does not observe θ directly, they must choose an action based on the data. A (randomized) decision rule $\delta : \mathcal{X} \to \Delta(\mathcal{A})$ maps data realizations to distributions over actions. A decision rule δ thus induces an expected loss that depends on (θ, X),

$$L_\delta(\theta, X) = \int L(a, \theta) d\delta(a; X).$$

We will take the induced loss functions $L : \Theta \times \mathcal{X} \to \mathbb{R}$ as our starting point and rank decision rules based on our preference over their induced losses L_δ. Our approach is thus similar in spirit to Stoye (2012), who takes risk (expected loss integrating over the distribution of the data) functions as a primitive to study the choice of decision rules. We work with loss rather than risk since risk depends on the distribution of the data, and we are interested in settings where this distribution may differ from that assumed by the model.

Example: Average Treatment Effects Suppose the researcher observes a sample of n observations $X = (X_1, ..., X_n)$ for $X_i = (Y_i, D_i)$, where $D_i \in \{0, 1\}$ is a binary treatment and $Y_i \in \{0, 1\}$ is a binary outcome. The unknown parameter $\theta = (\mu_0, \mu_1)$ collects the

mean potential outcomes $\mu_d = E[Y_i(d)]$ in a population of policy interest (e.g. where the researcher is considering rolling out the treatment), and the target parameter is the average treatment effect (ATE) $\kappa(\theta) = \mu_1 - \mu_0$. We consider loss $L(a, \theta) = (a - \kappa(\theta))^2$ and action space $\mathcal{A} = [-1, 1]$.

One decision rule the researcher could consider is the (non-randomized) difference-in-means estimator $\hat{\kappa}_{DM} = \bar{Y}_1 - \bar{Y}_0 = \frac{\sum_i D_i Y_i}{\sum_i D_i} - \frac{\sum_i (1-D_i) Y_i}{\sum_i 1-D_i}$, which induces loss $L_{DM}(\theta, X) = (\hat{\kappa}_{DM} - \kappa(\theta))^2$.[1] Another decision rule is the (randomized) half-sample difference in means $\hat{\kappa}_{DM,\frac{1}{2}}$, which drops half of the treatment and control observations, selected at random, and computes the difference in means over the remaining observations. This induces loss $L_{DM,\frac{1}{2}} = E\left[\left(\hat{\kappa}_{DM,\frac{1}{2}} - \kappa(\theta)\right)^2 | X\right]$ where the expectation is over the randomness induced by dropping observations. $\triangle$

Standard decision-theoretic analysis proceeds by first characterizing preferences over a large menu of options (e.g. loss functions) including ones which may be infeasible, and then applying these preferences to select from the feasible set for a particular problem. This approach is useful for characterizing preferences because the feasible set in a given setting is often quite restrictive, and it is easier to characterize preferences when working with a larger menu of choices. For instance, constant loss functions that assign the same loss regardless of the state are used in many decision theory results, but will often be infeasible in estimation problems: if we consider an estimation problem under squared error loss $L(a, \theta) = (a - \kappa(\theta))^2$, $L_\delta(\theta, X)$ will generally vary with θ so long as $\kappa(\theta)$ does.[2]

We assume the researcher has a statistical model $\mathcal{Q}$, which they may not fully trust. Formally, let $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$ denote the model, where each $Q_\theta \in \Delta(\mathcal{X})$ is a distribution for the data.[3] Let $\mathcal{L}$ denote the set of all bounded loss functions $L : \Theta \times \mathcal{X} \to \mathbb{R}$.[4] We take as primitive a family of preferences $\{\succsim_\Theta\}, \{\succsim_\theta : \theta \in \Theta\}$ over loss functions in $\mathcal{L}$. Throughout, $L \succsim L'$ means that $\succsim$ weakly prefers L to L'. The preference $\succsim_\Theta$ is the researcher's overall ranking of loss functions, and is the object we ultimately seek to characterize. The

[1] If either the treatment or control group is empty, define $\hat{\kappa}_{DM} = 0$.

[2] By contrast, Breza et al. (2025) consider an estimation problem where the researcher may decline to report an estimate at a constant cost, in which case certain constant loss functions are feasible.

[3] For simplicity, we further assume that $\mathcal{X}$ is Polish, that all distributions are defined on the Borel σ-algebra, and that functions discussed are Borel-measurable. Going forward, we suppress discussion of measure-theoretic details where possible.

[4] As is standard in the decision theory literature, proofs of the sufficiency direction of our representation theorems (the direction where the axioms imply the representation) will proceed on the domain of *simple* (finite-valued) loss functions. Under an appropriate continuity axiom, such a representation admits a unique extension to bounded loss functions.

auxiliary preferences $\{\succsim_\theta\}_{\theta\in\Theta}$ represent the researcher's evaluations given parameter value θ; we axiomatize these first and then derive $\succsim_\Theta$ from them. Intuitively, $\succsim_\theta$ represents the researcher's preference when they are certain that θ is the true parameter value but remain concerned that the data distribution Q_θ implied by their model may be misspecified.

By misspecification, we will mean that the data X are distributed according to P which differs from the distribution Q_θ implied by the true θ. Importantly, we assume the true value of θ remains well-defined even when the statistical model is wrong. This is natural for parameters like causal effects and counterfactuals which remain well-posed even when the model the researcher uses to estimate them is incorrect, but is less natural for e.g. parameters in a parametric utility, for which is it difficult to define a "true value" absent correct specification.[5] It also rules out the case where θ is defined as a statistical functional $\theta : \mathcal{Q} \to \mathbb{R}^p$, since in this case the model is either well-specified (when $P \in \mathcal{Q}$) or θ is undefined (when $P \notin \mathcal{Q}$). In the terminology of Andrews et al. (2025), we thus consider the problem of misspecification in the context of an *econometric model* described by the pair (θ, P), rather than a *statistical* model described solely by the distribution P of the data.

Example: Average Treatment Effects (continued) Suppose the researcher's model posits that the data are generated by (i) drawing a random sample from a population of interest which (ii) satisfies the potential outcomes model with each unit's outcome depending only on their own treatment and (iii) running a randomized trial where treatment is independently assigned to each unit with probability $\frac{1}{2}$. Under these assumptions $Y_i = Y_i(D_i)$, where D_i is independent of the potential outcomes $(Y_i(0), Y_i(1))$. Thus, under the researcher's model the observations $X_i = (Y_i, D_i)$ are i.i.d. draws from a multinomial distribution supported on $\{0,1\}^2$ with $E_{Q_\theta}[D_i] = \frac{1}{2}$, $E_{Q_\theta}[Y_i|D_i = 0] = \mu_0$, and $E_{Q_\theta}[Y_i|D_i = 1] = \mu_1$.

There are many ways in which the researcher's model could be wrong. For instance, the population from which experimental participants are sampled could differ from the population of policy interest (in which case the marginal distribution of potential outcomes $Y_i(d)$ in the trial could differ from that implied by θ). Alternatively, the experimental sample might not be drawn i.i.d., e.g. oversampling the friends of early experimental participants due to

[5] We could weaken this assumption to instead require only that some function $\kappa(\theta)$ that enters the loss, $L(a,\theta) = \tilde{L}(a,\kappa(\theta))$, have a model-agnostic definition, where the model implies a set of data distributions $\mathcal{Q}(\kappa^*) = \{Q_\theta : \theta \in \Theta, \kappa(\theta) = \kappa^*\}$ compatible with a given value κ^*. In particular, observe that for any function $f(Q_\theta)$, we have $\sup_{\theta\in\Theta} f(Q_\theta) = \sup_{\kappa^*} \sup_{Q\in\mathcal{Q}(\kappa)} f(Q_\theta)$, which may be used to reformulate our results in this format. While this is a weaker assumption on the interpretation of θ, it leads to substantially heavier exposition for some of our results, so we impose the stronger condition that the full vector θ has a model-free interpretation, or equivalently that κ does and that $\kappa(\cdot)$ is invertible.

recruitment through social networks (in which case the outcomes would not be independent across units). Finally, treatment might not be assigned as prescribed by the protocol, e.g. treating participants with lower baseline outcomes with higher probability (in which case treatment would not be independent of potential outcomes). If we combine these possibilities, any value of $\theta = (\mu_0, \mu_1)$ could in principle be compatible with any distribution $P \in \Delta(\{0,1\}^{2n})$ for the observable data X.

For the forms of misspecification discussed above, θ remains well-defined as the average potential outcomes in the target population, though it will not be identified absent restrictions on the possible misspecification. There are other forms of misspecification one could contemplate, e.g. spillovers across units, where the definition of θ becomes more delicate. If spillovers are present, for instance, does μ_0 correspond to the average outcome when no unit is treated, or when a given unit is untreated while others are treated i.i.d. with probability $\frac{1}{2}$? For our analysis, we presume the researcher has adopted a definition for the "true" θ which remains well-posed under the forms of misspecification they contemplate, and wishes to choose among decision rules in a way which is robust to this misspecification concern. $\triangle$

Results by Maccheroni et al. (2006) and Cerreia-Vioglio et al. (2025) imply axioms on the preference relations $\{\succsim_\Theta\}, \{\succsim_\theta : \theta \in \Theta\}$ which hold if and only if these preferences are represented by V_Θ and $\{V_\theta : \theta \in \Theta\}$ respectively, for

$$V_\Theta(L) = \sup_{\theta \in \Theta} \sup_{P \in \Delta(\mathcal{X})} \left\{ \int L(\theta, x) dP(x) - c_\theta(P) \right\} \tag{1}$$

$$V_\theta(L) = \sup_{P \in \Delta(\mathcal{X})} \left\{ \int L(\theta, x) dP(x) - c_\theta(P) \right\} \tag{2}$$

in the sense that

$$L \succsim_\Theta L' \iff V_\Theta(L) \leq V_\Theta(L')$$

$$L \succsim_\theta L' \iff V_\theta(L) \leq V_\theta(L').$$

In both representations, $c_\theta : \Delta(\mathcal{X}) \to [0, \infty]$ is a convex, lower-semicontinuous function with $c_\theta(Q_\theta) = 0$ for each $\theta \in \Theta$. Intuitively, the preference $\succsim_\theta$ evaluates risk functions L by (i) focusing on their behavior at parameter value θ and (ii) considering the penalized worst-case risk, which averages over P but then subtracts off $c_\theta(P)$, effectively penalizing expected loss under data distributions P that the researcher finds less plausible. We further assume that the researcher finds the model-implied distribution Q_θ at least as plausible as any other. The preference $\succsim_\Theta$ does the same for each θ, but further takes the worst case over all θ. As an

immediate consequence, V_Θ implies the (penalized) worst-case risk bound

$$\int L(\theta, x)dP(x) \leq V_\Theta(L) + c_\theta(P) \quad \forall \theta \in \Theta, P \in \Delta(\mathcal{X})$$

which controls, uniformly over (θ, P), how large the risk may be. Choosing a loss function which minimizes $V_\Theta(L)$ is thus the same as minimizing an upper bound on the risk.

Example: Average Treatment Effects, Continued Consider a researcher choosing between the Horvitz-Thompson estimator $\hat{\kappa}_{HT} = \frac{1}{n}\sum_i \frac{(2D_i-1)}{2}Y_i$ and the difference-in-means estimator $\hat{\kappa}_{DM} = \bar{Y}_1 - \bar{Y}_0$. Provided the researcher's preferences are represented by (1), they weakly prefer the difference-in-means estimator if and only if

$$V_\Theta(L_{DM}) = \sup_{\theta\in[0,1]^2} \sup_{P\in\Delta(\{0,1\}^{2n})} \int \left(\hat{\kappa}_{DM} - \kappa(\theta)\right)^2 dP(x) - c_\theta(P) \leq$$

$$\sup_{\theta\in[0,1]^2} \sup_{P\in\Delta(\{0,1\}^{2n})} \int \left(\hat{\kappa}_{HT} - \kappa(\theta)\right)^2 dP(x) - c_\theta(P) = V_\Theta(L_{HT}).$$

Moreover, the optimized value $V_\Theta(L_{DM})$ of the left hand side provides a bound on how quickly the performance of the difference in means estimator moves away from that prescribed by the researcher's model. $\triangle$

The representations (1) and (2) follow from arguments in Maccheroni et al. (2006) and Cerreia-Vioglio et al. (2025), but our framing of the problem (e.g. choice domain, preferences) differs from theirs. For completeness, and to aid interpretation for readers more used to working with risk and loss functions than the conventional decision-theory setup, we thus provide an axiomatization for our choice domain, along with proofs, in Appendix A.

The axiomatization discussed in Appendix A leaves the form of the penalty function c unspecified. Different choices of c correspond to different attitudes toward misspecification, and restricting these attitudes narrows the class of penalties. We next discuss three important cases: constraint preferences, multiplier preferences, and a novel class of constrained multiplier preferences that combines the other two. In each case, we discuss the additional axioms that distinguish these preferences from the broader class characterized (1) and (2).

2.1 Constraint Preferences

A particularly simple penalty arises when the researcher requires that the true distribution P lie in some set but imposes no further penalties. Consider a closed, convex set $\mathcal{P}_\theta \subseteq \Delta(\mathcal{X})$ of distributions (the "ambiguity set") that the researcher treats as plausible under parameter value θ. We require that $Q_\theta \in \mathcal{P}_\theta$ so that the researcher finds their model plausible. Preferences of the form

$$V_\Theta(L) = \sup_{\theta\in\Theta} \sup_{P\in\mathcal{P}_\theta} \int L(\theta, x) dP(x)$$

$$V_\theta(L) = \sup_{P\in\mathcal{P}_\theta} \int L(\theta, x) dP(x) \tag{3}$$

evaluate loss functions based on their worst-case performance over both parameters θ and distributions P in the (θ-specific) ambiguity set. Criteria of this form were recently discussed in Bonhomme and Weidner (2022), while similar misspecification-aware setups are considered for other purposes (e.g. confidence set construction) by Armstrong and Kolesár (2021) and Christensen and Connault (2023).[6]

Example: Average Treatment Effects, Continued Now suppose the experiment is conducted by two teams, with unit i randomly assigned to team $C_i \in \{1, 2\}$, where $\pi_j = P\{C_i = j\}$. The researcher's model maintains that outcomes do not depend on team assignment, so $E[Y_i(d) \mid C_i = j] = \mu_d$ for all j, and the likelihood pools data from both teams. However, the researcher has greater confidence in team 1's execution than in team 2's: for instance, team 2 may have deviated from the treatment assignment protocol, implemented the treatment less carefully, or measured outcomes differently. We may express this using constraint preferences.

Let $\mathcal{P}_\theta$ be the set of distributions which arise by (i) drawing C_i i.i.d. with $P\{C_i = j\} = \pi_j$ (ii) drawing $(Y_i, D_i)|C_i = 1$ as described by the researcher's model and (iii) drawing the remaining outcomes $\{(Y_i, D_i)|i \in \{1, ..., n\}$ such that $C_i = 2\}$ from some other distribution. This case lies strictly between full trust in the model ($\mathcal{P}_\theta = \{Q_\theta\}$) and complete agnosticism ($\mathcal{P}_\theta = \Delta(\mathcal{X}^n)$): it requires that team 1's observations follow the model, while placing no restrictions (including independence) on team 2's observations. $\triangle$

[6]Note that while the class $\mathcal{P}_\theta$ considered in e.g. Bonhomme and Weidner (2022) need not be convex, $V_\Theta(L)$ and $V_\theta(L)$ are unchanged if we replace all non-convex $\mathcal{P}_\theta$ by their convex hull.

The preference (3) is special case of (2) with the penalty function

$$c_\theta(P) = \begin{cases} 0 & \text{if } P \in \mathcal{P}_\theta \\ \infty & \text{if } P \notin \mathcal{P}_\theta \end{cases}. \tag{4}$$

These preferences, which can be viewed as a version of Gilboa and Schmeidler (1989) min-max preferences, are distinguished from the more general class of variational preferences by the Certainty Independence axiom. Certainty independence is stated using a *constant loss*, by which we mean a loss function $r \in \mathcal{L}$ that takes the same value $r \in \mathbb{R}$ for all (θ, x), and we slightly abuse notation by identifying constant functions with their value.

Axiom 1 (Certainty Independence)**.** *For all $\theta \in \Theta$, $L, L' \in \mathcal{L}$, $r \in \mathbb{R}$, and $\alpha \in (0,1)$,*

$$L \succsim_\theta L' \iff \alpha L + (1-\alpha)r \succsim_\theta \alpha L' + (1-\alpha)r.$$

Intuitively, certainty independence requires that our preference over estimators not change if, with some probability independent of the data and parameter, we will be randomly switched to instead use an estimator with constant loss.

Example: Average Treatment Effects, Continued Again consider a researcher choosing between Horvitz-Thompson and difference-in-means. Now suppose that for each data realization this researcher is, with probability $1-\alpha$, randomized into instead using a noisy oracle $\hat{\kappa}_O = \kappa(\theta) + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. This results in loss functions $\alpha L_{HT} + (1-\alpha)\sigma^2$ and $\alpha L_{DM} + (1-\alpha)\sigma^2$, and Certainty Independence requires that for all $\alpha \in (0,1)$, the researcher's preference between $\hat{\kappa}_{HT}$ and $\hat{\kappa}_{DM}$ be the same as in the problem without randomization (i.e. with $\alpha = 1$). $\triangle$

Proposition 1 (Gilboa and Schmeidler, 1989)**.** *Suppose that the conditional preferences $\{\succsim_\theta : \theta \in \Theta\}$ are represented by (2). Certainty Independence holds if and only if $c_\theta(\cdot)$ is of the form (4) for all $\theta \in \Theta$.*

While Proposition 1 characterizes when the researcher's preferences have an (unpenalized) min-max form, the class of possible ambiguity sets $\mathcal{P}_\theta$, and thus preferences, remains quite large. Ghirardato and Marinacci (2002) show that one can go further and infer the ambiguity sets $\mathcal{P}_\theta$ directly from $\succsim_\theta$.

To state this result, we must introduce some additional notation. For $(\theta, P) \in \Theta \times \Delta(\mathcal{X})$, define the *subjective expected utility (SEU) preference* $\succsim^{SEU}_{\theta,P}$ by

$$L \succsim^{SEU}_{\theta,P} L' \iff \int L(\theta, x)dP(x) \leq \int L'(\theta, x)dP(x).$$

This is the preference of a decision maker who ranks decision rules based on their expected loss under (θ, P). Following Ghirardato and Marinacci (2002), $\succsim_\theta$ is *more ambiguity averse* than $\succsim^{SEU}_{\theta,P}$ if for all $L \in \mathcal{L}$ and $r \in \mathbb{R}$,

$$L \succsim_\theta r \Rightarrow L \succsim^{SEU}_{\theta,P} r.$$

In words, any time the preference $\succsim_\theta$ ranks a state-dependent loss L as weakly better than a constant loss r, $\succsim_{\theta,P}$ must do so as well. Equivalently, $\succsim_\theta$ is "more cautious" than expected loss under (θ, P): it has a (weakly) higher bar for preferring uncertain losses to certain ones. Ghirardato and Marinacci (2002) establish that such comparisons identify $\mathcal{P}_\theta$.

Proposition 2 (Ghirardato and Marinacci, 2002)**.** *Suppose the conditions of Proposition 1 hold, and let*

$$\mathcal{P}_\theta = \{P \in \Delta(\mathcal{X}) : \succsim_\theta \text{ is more ambiguity averse than } \succsim^{SEU}_{\theta,P}\}.$$

Then $c_\theta(\cdot)$ is equal to (4) *with this $\mathcal{P}_\theta$.*

In particular, since we have assumed that $Q_\theta \in \mathcal{P}_\theta$, Proposition 2 implies that $\succsim_\theta$ is more ambiguity averse than $\succsim_{\theta,Q_\theta}$.

Example: Average Treatment Effects, Continued Consider a researcher choosing between the difference-in-means estimator $\hat{\kappa}_{DM} = \bar{Y}_1 - \bar{Y}_0$ and the noisy oracle $\hat{\kappa}_O = \kappa(\theta) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$. If the researcher's conditional preferences $\succsim_\theta$ are represented by (2) and satisfy Certainty Independence, and thus have a constraint representation with $Q_\theta \in \mathcal{P}_\theta$ for each $\theta \in \Theta$, it follows that $L_O \succsim_\theta L_{DM}$ whenever $\sigma^2 \leq E_{Q_\theta}[(\hat{\kappa}_{DM} - \kappa(\theta))^2]$. $\triangle$

2.2 Multiplier Preferences

Constraint preferences compute the worst-case expected loss over $\mathcal{P}_\theta$, and thus do not privilege any particular distribution in this set. Since we have assumed the researcher has a

model $\mathcal{Q}$, however, it also seems natural that under parameter value θ they might find distributions "close" to Q_θ more plausible than those "far" from Q_θ, and thus wish to penalize some notion of distance from the base model. As has previously been observed many times (e.g. by Hansen and Sargent, 2001, 2008), Kullback-Leibler (KL) divergence is an especially convenient penalty for many purposes, and leads to

$$V_\Theta(L) = \sup_{\theta\in\Theta} \sup_{P\in\Delta(\mathcal{X})} \left\{ \int L(\theta, x) dP(x) - \lambda \cdot \mathrm{KL}(P\|Q_\theta) \right\},$$

$$V_\theta(L) = \sup_{P\in\Delta(\mathcal{X})} \left\{ \int L(\theta, x) dP(x) - \lambda \cdot \mathrm{KL}(P\|Q_\theta) \right\}, \tag{5}$$

where $\lambda > 0$ and

$$\mathrm{KL}(P\|Q) = \int \log\left(\frac{dP}{dQ}(x)\right) dP(x)$$

when $P \ll Q$ (i.e. all events which are probability zero under Q are also probability zero under P) and $\mathrm{KL}(P\|Q) = \infty$ otherwise. The parameter λ controls the degree of concern for misspecification: as $\lambda \to \infty$, the preference converges to expected loss under Q_θ, while as $\lambda \to 0$, the preference places more weight on worst-case scenarios.

Results in the literature again characterize this penalty relative to the broad class $c_\theta(\cdot)$. Strzalecki (2011) shows a tight connection between the KL divergence and Savage's Sure Thing Principle. To state the Sure Thing Principle, for all $L, L' \in \mathcal{L}$ define a spliced loss equal to L on an event $\mathcal{E}$ and L' otherwise,

$$L_\mathcal{E} L'(\theta, X) = \begin{cases} L(\theta, X) & \text{if } X \in \mathcal{E} \\ L'(\theta, X) & \text{if } X \notin \mathcal{E} \end{cases}.$$

Axiom 2 (Sure Thing Principle)**.** *For all $\theta \in \Theta$, $\mathcal{E} \subseteq \mathcal{X}$, and $L, L', M, M' \in \mathcal{L}$,*

$$L_\mathcal{E} M \succsim_\theta L'_\mathcal{E} M \iff L_\mathcal{E} M' \succsim_\theta L'_\mathcal{E} M'.$$

The Sure Thing Principle requires that preferences between loss functions that agree on some set ($\mathcal{E}^c$, in this case) depend only on their values elsewhere. In our setting, this axiom has a natural connection to pretesting.

Example: Average Treatment Effects, Continued The experimental design implies $E_{Q_\theta}[D_i] = \frac{1}{2}$, which the researcher could test in order to check implementation fidelity. Suppose that, conditional on not rejecting the null $H_0 : E_P[D_i - \frac{1}{2}] = 0$, the researcher will choose between the Horvitz-Thompson estimator $\hat{\kappa}_{HT}$ and the difference-in-means estimator $\hat{\kappa}_{DM}$, while if the null is rejected they will use some alternative procedure. The Sure Thing Principle requires that the ranking between $\hat{\kappa}_{HT}$ and $\hat{\kappa}_{DM}$ under $\succsim_\theta$, conditional on the test not rejecting, not vary depending on what procedure is used when the test rejects. $\triangle$

The results of Strzalecki (2011) imply that the Sure Thing Principle holds if and only if $c_\theta(\cdot)$ is KL divergence relative to some centering distribution. Our requirement that $c_\theta(Q_\theta) = 0$ further ensures that this centering distribution is Q_θ (see Appendix A for an axiomatic justification). We further impose an axiom, adapted from Lanzani (2025), which enforces that the misspecification-aversion parameter λ is constant across θ.

Axiom 3 (Uniform Misspecification Concern)**.** *For each $\theta, \theta' \in \Theta$ and each $L, L' \in \mathcal{L}$ with*

$$Q_\theta \circ L_\theta^{-1} = Q_{\theta'} \circ (L'_{\theta'})^{-1} \tag{6}$$

where $Q_\theta \circ L_\theta^{-1}$ denotes the distribution of $L(\theta, X)$ when $X \sim Q_\theta$, it holds that

$$L \succsim_\theta r \iff L' \succsim_{\theta'} r \quad \forall r \in \mathbb{R}$$

Example: Average Treatment Effects, Continued Suppose the researcher is considering an estimator $\hat{\kappa}$ whose induced distribution over squared estimation errors $(\hat{\kappa} - \kappa(\theta))^2$ does not depend on θ. For example, consider the (infeasible) estimator $\hat{\kappa}$ which yields an estimation error of $1/2$ if the sample fraction of treated individuals exceeds $1/2$ and 0 otherwise.[7] Since the law of the sample fraction of treated individuals is $n^{-1}\text{Bin}(n, 1/2)$ under Q_θ for all θ, $\hat{\kappa}$ satisfies the property (6) for all $\theta, \theta' \in \Theta$. $\triangle$

Axiom 4 (Monotone Continuity)**.** *For all $\theta \in \Theta$, $L, L' \in \mathcal{L}$, $r \in \mathbb{R}$, and sequences of events*[8] *$\{\mathcal{E}_n\}_{n\geq 1}$ with $\mathcal{X} \supseteq \mathcal{E}_1 \supseteq \mathcal{E}_2 \supseteq \cdots$ and $\cap_{n\geq 1}\mathcal{E}_n = \emptyset$: if $L \succ_\theta L'$, there exists $n^* \geq 1$ such that $r_{\mathcal{E}_{n^*}} L \succ_\theta L'$.*

Loosely speaking, Monotone Continuity ensures that the representation only cares about countably additive probabilities. When $\mathcal{X}$ is finite, it has no bite.

[7] Note that in the limit experiment discussed in the next section, there do exist estimators whose model-implied distribution does not depend on the parameter.

[8] In this axiom, events are Borel subsets of $\mathcal{X}$.

Finally, say that an event $\mathcal{E} \subseteq \mathcal{X}$ is *nonnull* under $\succsim_\theta$ if there exist $L, L', M \in \mathcal{L}$ such that $L_\mathcal{E} M \succ_\theta L'_\mathcal{E} M$. To state our representation theorem, we make the mild assumption that $\mathcal{X}$ has at least three disjoint nonnull events for each $\succsim_\theta$, and that there exists $q \in (0,1)$ such that for each θ, there exists an event $\mathcal{E}_\theta \subseteq \mathcal{X}$ with $Q_\theta(\mathcal{E}_\theta) = q$.

Proposition 3 (Strzalecki 2011, Cerria-Vioglio et al. 2025)**.** *Suppose that the conditional preferences* $\{\succsim_\theta : \theta \in \Theta\}$ *are represented by* (2)*. The Sure Thing Principle, Uniform Misspecification Concern, and Monotone Continuity hold if and only if* $c_\theta(P) = \lambda \cdot \mathrm{KL}(P \| Q_\theta)$ *for some* $\lambda > 0$*.*

2.3 Constrained Multiplier Preferences

Multiplier preferences treat all forms of misspecification symmetrically: deviations from the model are penalized solely based on their KL divergence. This is natural when the researcher has no view about which aspects of the model are more likely to fail. In many economic applications, however, researchers appear to have more confidence in some model predictions than in others. For such researchers, it is natural to combine both elements: a hard constraint ruling out certain forms of misspecification a-priori, and a KL penalty governing concern about remaining misspecification within the ambiguity set.

To capture the resulting preferences, we introduce a novel class of *constrained multiplier preferences*, which rule out some DGPs a-priori and then continuously penalize deviations deviations within the ambiguity set using Kullback-Leibler divergence:

$$V_\Theta(L) = \sup_\theta \sup_{P \in \mathcal{P}_\theta} \left\{ \int L(\theta, x) dP(x) - \lambda \cdot \mathrm{KL}(P \| Q_\theta) \right\} \tag{7}$$

$$V_\theta(L) = \sup_{P \in \mathcal{P}_\theta} \left\{ \int L(\theta, x) dP(x) - \lambda \cdot \mathrm{KL}(P \| Q_\theta) \right\}. \tag{8}$$

Mathematically, (7) and (8) correspond to special cases of (1) and (2), respectively, which take $c_\theta(\cdot)$ equal to the sum of the convex indicator for the set $\mathcal{P}_\theta$, as in constraint preferences, and the KL divergence from the model distribution Q_θ, as in variational preferences.

Example: Average Treatment Effects, Continued In our discussion of constraint preferences we considered the case where $\mathcal{P}_\theta$ exactly pins down the distribution of the data collected by team 1 while imposing no constraints on the data from team 2. This represents extreme distrust of the second team, for instance treating it as equally plausible that this team ad-

hered faithfully to the experimental protocol and that they fabricated the data wholesale. Constrained multiplier preferences accommodate the intermediate case where the researcher thinks the data from the second team may have a distribution different than that predicted by the model (in principle allowing any $P \in \mathcal{P}_\theta$ with $KL(P\|Q_\theta) < \infty$) but continuously discounts distributions P which are further from Q_θ as measured by KL divergence. $\triangle$

Our axiomatization of constrained multiplier preferences will build on the constraint and multiplier preference axiomizations. Specifically, let $\succsim_\theta^C$ denote the preference the researcher would have if they (i) took as given that the true parameter is θ and (ii) were certain the true distribution lay in $\mathcal{P}_\theta$ but did not privilege any distribution in this set. Formally, we assume the preferences $\left\{\succsim_\theta^C : \theta \in \Theta\right\}$ satisfy the conditions of Propositions 1 and 2, and thus are constraint preferences with ambiguity set $\mathcal{P}_\theta$.

Similarly, let $\succsim_\theta^M$ denote the preference the researcher would have if they (i) took as given that the true parameter is θ and (ii) were more concerned with performance "close" to the model-implied distribution Q_θ but did not have hard constraints on the class of possible data distributions and (iii) satisfy the sure thing principle. Formally, we assume the preferences $\left\{\succsim_\theta^M : \theta \in \Theta\right\}$ satisfy conditions of Proposition 3, and thus are multiplier preferences with centering distribution Q_θ.

We consider a researcher who both believes the constraints on misspecification imposed by $\succsim_\theta^C$ and is more concerned with DGPs close to Q_θ as in $\succsim_\theta^M$. The next axioms captures how their conditional preference given θ, $\succsim_\theta$, combines these two elements.

Axiom 5 (Indirect Pareto)**.** *For all $\theta \in \Theta$ $L \in \mathcal{L}$ and $r \in \mathbb{R}$,*

$$L \succ_\theta r \iff$$

$$\exists\ L_C, L_M \in \mathcal{L},\ r_C, r_M \in \mathbb{R}\ s.t.\ L = L_C + L_M,\ r = r_C + r_M,\ L_C \succ_\theta^C r_C,\ L_M \succ_\theta^M r_M.$$

The Indirect Pareto axiom requires that the conditional preference $\succsim_\theta$ strictly prefers the loss function L to a constant loss r if and only if L can be decomposed into two parts, corresponding to the constraint and multiplier preferences respectively, each of which is strictly preferred to its share of the constant under the respective component preference. We show that this property characterizes the constrained multiplied preference.

Theorem 1. *Suppose each $\succsim_\theta^C$ is a constraint preference with ambiguity set $\mathcal{P}_\theta$ and each $\succsim_\theta^M$ is a multiplier preference with centering measure Q_θ and parameter λ. Under Axiom 5, the preference $\succsim_\theta$ has the constrained multiplier representation (8).*

2.4 Dual Representation of Constrained Multiplier Preferences

One practically appealing feature of multiplier preferences is that they imply highly tractable dual representations.

Proposition 4 (Dupis and Ellis, 1997)**.** *For V_θ as in 5,*

$$V_\theta(L) = \lambda \cdot \log\left(E_{Q_\theta}\left[\exp\left(\frac{1}{\lambda}L(\theta, X)\right)\right]\right). \tag{9}$$

This result shows that the ranking over losses implied by multiplier preferences is precisely the same as the ranking one would obtain by assuming the researcher's model Q_θ is correct but using the exponentiated loss $\exp\left(\frac{1}{\lambda}L\right)$. Consequently, one can compute optimal decision rules under multiplier preferences by applying standard arguments for the correctly-specified case to the transformed loss function.

Parallel convex duality arguments also imply a tractable dual for constrained multiplier preferences in some contexts. In particular, we focus on the case where the ambiguity set $\mathcal{P}_\theta$ can be written as the set of distributions satisfying a collection of moment equalities. For ambiguity sets of this form, convex duality arguments similar to those in the generalized empirical likelihood literature (Newey and Smith, 2004; Kitamura, 2009) imply a finite-dimensional dual for the constrained multiplier problem.

Proposition 5. *Suppose $\mathcal{P}_\theta = \{P \in \Delta(\mathcal{X}) : E_P[\varphi(\theta, X)] = 0\}$ for a vector of moment functions $\varphi : \mathcal{X} \times \Theta \to \mathbb{R}^b$ satisfying $E_{Q_\theta}[\varphi(\theta, X)] = 0$. Then for V_θ as in (8),*

$$V_\theta(L) = \inf_{\beta \in \mathbb{R}^b} \lambda \cdot \log\left(E_{Q_\theta}\left[\exp\left(\frac{1}{\lambda}L(\theta, X) - \beta'\varphi(\theta, X)\right)\right]\right). \tag{10}$$

The representation (10) introduces Lagrange multipliers β that enforce the moment constraints. The infimum over β is the dual to the original problem of maximizing over $\mathcal{P}_\theta$. While this optimization problem does not in general have a closed form solution, it is convex and sufficiently tractable to enable both computation and theoretical analysis.

Example: Average Treatment Effects, Continued Consider the two-team ATE setting introduced above. The ambiguity set $\mathcal{P}_\theta$ restricts the joint distribution of the team-1 observations. Fully expressing these constraints using moment equalities would require per-observation restrictions (e.g. $E_P[(Y_i - \mu_0)(1 - D_i)1\{C_i = 1\}] = 0$ for each i) and cross-observation restrictions (e.g. zero covariance between $(D_i - \frac{1}{2})1\{C_i = 1\}$ and $(D_j - \frac{1}{2})1(C_j =$

1) for $i \neq j$), with the total number of moment equalities growing with the sample size.

To obtain a more parsimonious (but also more permissive) set of moment conditions, we may instead constrain a sample average moment function. Let

$$\psi(\theta, X_i) = \begin{pmatrix} (Y_i - \mu_0)(1 - D_i)1\{C_i = 1\} \\ (Y_i - \mu_1)D_i 1\{C_i = 1\} \\ (D_i - \frac{1}{2})1\{C_i = 1\} \\ 1\{C_i = 1\} - \pi_1 \end{pmatrix} \tag{11}$$

and define $\varphi(\theta, X) = \frac{1}{n}\sum_{i=1}^{n} \psi(\theta, X_i)$. The first two components capture the conditional mean outcomes among treated and control units in team 1, while the third and fourth elements restrict treatment and team assignment, respectively. The constraint $E_P[\varphi(\theta, X)] = 0$ requires that the moment conditions hold on average across units, but does not constrain the marginal for a given unit and thus leads to a weakly higher worst-case risk. As we discuss in the next section we can strengthen the constraint by including higher moments of the sample average.

With the moment-equality constraint set $\mathcal{P}_\theta = \{P : E_P[\varphi(\theta, X)] = 0\}$, Proposition 5 applies, and the researcher's preference $\succsim_\Theta$ ranks estimators $\hat{\kappa}_\delta$ based on

$$\sup_{\theta \in \Theta} \inf_{\beta \in \mathbb{R}^4} \lambda \cdot \log \left(E_{Q_{n,\theta}} \left[\exp \left(\frac{1}{\lambda} (\hat{\kappa}_\delta - \kappa(\theta))^2 - \beta' \varphi(\theta, X) \right) \right] \right),$$

where we have used that the ranking is unchanged by monotone transformations of V_Θ. $\triangle$

Before moving on, we briefly note that the analytic tractability of constrained multiplier preferences extends to settings with moment inequality, rather than equality, constraints. Specifically, if $\mathcal{P}_\theta = \{P : E_P[\varphi(\theta, X)] \leq 0\}$, one can extend Proposition 5 to show that

$$V_\theta(L) = \inf_{\beta \in \mathbb{R}^b} \sup_{\eta \in \mathbb{R}^b_+} \lambda \cdot \log \left(E_{Q_\theta} \left[\exp \left(\frac{1}{\lambda} L(\theta, X) - \beta' (\varphi(\theta, X) + \eta) \right) \right] \right) =$$

$$\inf_{\beta \in \mathbb{R}^b_+} \lambda \cdot \log \left(E_{Q_\theta} \left[\exp \left(\frac{1}{\lambda} L(\theta, X) - \beta' (\varphi(\theta, X)) \right) \right] \right).$$

3 Asymptotic Risk Bounds

While we derived constrained multiplier preferences in a finite-sample setting, in most interesting economic models finite-sample performance is analytically intractable. Following a

foundational analytic approach for models without misspecification concern, we thus study local asymptotic performance instead. This section develops a local asymptotic minimax (LAM) theorem for constrained multiplier preferences, paralleling a classical result for the correctly specified case. We begin by introducing the asymptotic framework and reviewing the classical LAM theorem as a benchmark, then state our LAM result and discuss its implications.

3.1 The Classical LAM Theorem

Consider a sequence of estimation problems indexed by the sample size n. For sample size n the researcher observes data $X^n = (X_1, \ldots, X_n)$, which under their model is drawn from a distribution in $\mathcal{Q}_n = \{Q_{n,\theta} : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^p$. For instance, if the researcher's model implies the data are i.i.d. then $Q_{n,\theta} = \times_{i=1}^n Q_{1,\theta}$ for $Q_{1,\theta}$ the distribution of a single observation, though the framework allows for more general dependence structures. We study performance when the true parameter is local to a base value θ_0, taking the form $\theta_{n,h} = \theta_0 + h/\sqrt{n}$ for a local parameter $h \in H = \mathbb{R}^p$, and we shorthand $Q_{n,\theta_{n,h}} = Q_{n,h}$. The loss function in the sample of size n is

$$L_n(a, \theta) = \ell(\sqrt{n}(a - \kappa(\theta))),$$

where $\kappa : \Theta \to \mathbb{R}^d$ is the target parameter and $\ell : \mathbb{R}^d \to \mathbb{R}$ is a fixed loss function. We are interested in the worst-case asymptotic performance of decision rule sequences δ_n over the local parameter space H.

The notion of local asymptotic normality, a foundational tool for characterizing asymptotic performance, formalizes a sense in which regular statistical models are asymptotically equivalent to Gaussian location experiments.

Definition 1. *The sequence of models $\{Q_{n,\theta} : \theta \in \Theta\}$ is* locally asymptotically normal *(LAN) at θ_0 with scaling coefficient $\sqrt{n}$ if there exists a sequence of random vectors S_{n,θ_0} and a nonsingular matrix I_0 such that for every sequence $h_n \to h$,*

$$\log\left(\frac{dQ_{n,\theta_0+h_n/\sqrt{n}}}{dQ_{n,\theta_0}}\right) = h^T S_{n,\theta_0} - \frac{1}{2} h^T I_0 h + o_{Q_{n,\theta_0}}(1),$$

where $S_{n,\theta_0} \to_d N(0, I_0)$ under Q_{n,θ_0}.

The LAN condition says that, in a local neighborhood of θ_0, the log-likelihood ratio is

asymptotically quadratic in the local parameter h with Hessian $-I_0$, and thus resembles the log-likelihood of a normal model with Fisher Information (and inverse variance) I_0. For i.i.d. data, LAN follows from standard differentiability conditions on the single-observation density (see e.g. van der Vaart, 1998, Chapter 7).

For LAN models, the classical local asymptotic minimax theorem gives a lower bound on the worst-case local asymptotic risk of any sequence of estimators.

Proposition 6 (Local Asymptotic Minimax; van der Vaart 1998, Theorems 8.11 and 9.4)**.** *Suppose $\{Q_{n,\theta} : \theta \in \Theta\}$ is LAN at $\theta_0 \in \text{int}(\Theta)$ with nonsingular Fisher information I_0. Then for any symmetric, quasi-convex loss $\ell : \mathbb{R}^d \to \mathbb{R}_+$ minimized at zero, any $\kappa : \Theta \to \mathbb{R}^d$ with derivative $K = \frac{\partial}{\partial \theta'}\kappa(\theta_0) \in \mathbb{R}^{d \times p}$, and any sequence of decision rules δ_n,*

$$\sup_I \liminf_{n\to\infty} \sup_{h \in I} E_{Q_{n,h}}\left[\ell\left(\sqrt{n}(\delta_n(X^n) - \kappa(\theta_{n,h}))\right)\right] \geq \int \ell \, dN(0, KI_0^{-1}K'),$$

where the supremum ranges over finite subsets I of $H = \mathbb{R}^p$.

The right-hand side is the minimax risk in the Gaussian limit experiment where one observes $X \sim N(h, I_0^{-1})$ and wishes to estimate Kh under loss ℓ. The bound says no sequence of estimators can achieve lower worst-case local asymptotic risk then the finite-sample risk in this Gaussian problem. Under squared error loss, the bound reduces to $KI_0^{-1}K'$, achieved by any efficient estimator, including the maximum likelihood estimator. The classical LAM theorem thus provides a theoretical foundation for familiar efficiency claims.

Example: Average Treatment Effects, Continued Under the model described in Section 2, the family $\{Q_{n,\theta}\}$ is LAN at any interior $\theta_0 = (\mu_0, \mu_1)$ with Fisher information $I_0 = \frac{1}{2}\,\text{diag}(1/\sigma_0^2, 1/\sigma_1^2)$ for $\sigma_d^2 = \mu_d(1-\mu_d)$. Since $K = (-1, 1)$, the classical LAM bound under squared error loss is $KI_0^{-1}K' = 2(\sigma_0^2 + \sigma_1^2)$, achieved by the difference-in-means estimator. $\triangle$

3.2 The Constrained Multiplier LAM Theorem

We now develop an analogous result for constrained multiplier preferences. The key additional ingredient is a moment function $\psi : \Theta \times \mathcal{X} \to \mathbb{R}^k$ satisfying $E_{Q_{n,\theta}}[\psi(\theta, X_i)] = 0$ for all n and θ. This moment function encodes the researcher's beliefs about which aspects of the model are correctly specified. We assume that the researcher believes misspecification does not affect the first M moments of the scaled sample average of ψ evaluated at θ_0. Including

higher moments is potentially important, since the forms of misspecification allowed by constrained multiplier preferences include ones which change the dependence structure of the data. Thus, even if misspecification does not affect the marginal distribution of each observation, the distribution of sample averages could still change. Constraining the moments of the sample average up to order M restricts such possibilities.

Formally, let

$$Y_{n,h} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\theta_{n,h}, X_i),$$

and for $m = (m_1, \ldots, m_k) \in \mathbb{N}_0^k$ with $1 \leq \sum_{s=1}^k m_s \leq M$, define $\tilde{W}_{n,h}^m = \prod_{s=1}^k Y_{n,h,s}^{m_s}$. Let $\tilde{W}_{M,n,h}$ collect $\tilde{W}_{n,h}^m$ over all such m, and define

$$W_{M,n,h} = \tilde{W}_{M,n,h} - E_{Q_{n,h}}[\tilde{W}_{M,n,h}].$$

The constraint set

$$\mathcal{P}_{n,h}^M = \{P \in \Delta(\mathcal{X}^n) : E_P[W_{M,n,h}] = 0\}$$

consists of all data distributions that preserve the first M moments of $Y_{n,h}$. This takes the form assumed in Proposition 5 with $\varphi(\theta_{n,h}, X^n) = W_{M,n,h}$, so the duality result applies.

Example: Average Treatment Effects, Continued Returning to the two-team ATE example, recall the per-observation moment function (11). The constraint set $\mathcal{P}_{n,h}^M$ enforces that misspecification not affect the first M moments of $Y_{n,h} = \frac{1}{\sqrt{n}} \sum_i \psi(\theta_{n,h}, X_i) \in \mathbb{R}^4$. For $M = 1$, as discussed above this constrains the mean of $Y_{n,h}$, requiring that $\sum_{i=1}^n E_P[\psi(\theta_{n,h}, X_i)] = 0$. For $M \geq 2$, the constraints additionally involve moments of $Y_{n,h}$ that depend on pairwise covariances of $\psi(\theta, X_i)$ across observations, restricting the impact of misspecification of the cross-observation dependence within team 1's data. $\triangle$

To derive our LAM theorem we impose two assumptions. The first collects regularity conditions on the model and the moment function.

Assumption 1. *The family $\{Q_{n,\theta} : \theta \in \Theta\}$ is stationary for each n with*

$$E_{Q_{n,\theta}}[\psi(\theta, X_i)] = 0 \quad \textit{for all } \theta \in \Theta, \qquad E_{Q_{n,\theta_0}}\left[\frac{\partial}{\partial \theta'} \psi(\theta_0, X_i)\right] = \Psi.$$

Moreover, $\{Q_{n,\theta} : \theta \in \Theta\}$ *has densities* $\{q_{n,\theta} : \theta \in \Theta\}$*, where for*

$$S_n = \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \log q_{n,\theta_0}(X^n), \qquad Y_{n,0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\theta_0, X_i),$$

$$\begin{pmatrix} S_n \\ Y_{n,0} \end{pmatrix} \overset{Q_{n,\theta_0}}{\underset{d}{\to}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_0 & -\Psi' \\ -\Psi & \Omega \end{pmatrix} \right)$$

where Ω *has full rank. Finally,* $\psi(\theta, X_i)$ *and* $q_{n,\theta}(X^n)$ *are differentiable at* θ_0 *for all* X_i *and* X^n*, and there exist an open neighborhood* $\mathcal{N}$ *of* θ_0 *and a constant* $C < \infty$ *such that*

$$E_{Q_{n,\theta_0}} \left[\sup_{\theta \in \mathcal{N}} \left(\left\| \frac{\partial}{\partial \theta} \psi(\theta, X_i) \right\| + \left\| \frac{\partial}{\partial \theta} \log q_{n,\theta}(X^n) \right\| \right) \right] \leq C$$

for all n*.*

These conditions are standard: stationarity, a central limit theorem for the scaled moments, and smoothness of the moment function and likelihood. We also restrict the loss.

Assumption 2. *The loss function in the sample of size* n *is equal to* $L_n(a, \theta) = \ell(\sqrt{n}(a - \kappa(\theta)))$*, where* $\kappa : \Theta \to \mathbb{R}^d$ *is differentiable at* θ_0 *and* $\ell : \mathbb{R}^d \to [0, \infty)$ *is convex, finite-valued, and satisfies* $\ell(u) \to \infty$ *as* $\|u\| \to \infty$*.*

Relative to the classical LAM theorem, Assumption 2 strengthens quasi-convexity to full convexity, but allows asymmetric loss. Convexity ensures that randomized estimators cannot improve on deterministic ones, which simplifies our asymptotic results.

By Proposition 5, in the sample of size n the worst-case constrained multiplier risk of an estimator δ_n over a set of local parameters $I \subset H$ is

$$\sup_{h \in I} \sup_{P \in \mathcal{P}^M_{n,h}} \{\mathbb{E}_P [L_n(\delta_n(X^n), \theta_{n,h})] - \lambda D_{KL}(P \| Q_{n,h})\} =$$

$$\sup_{h \in I} \inf_{\beta} \lambda \cdot \log \left(E_{Q_{n,h}} \left[\ell^* \left(\sqrt{n} \left(\delta_n(X^n) - \kappa(\theta_{n,h}) \right) \right) \exp \left(\beta' W_{M,n,h} \right) \right] \right)$$

for $\ell^*(u) = \exp\left(\frac{1}{\lambda} \ell(u)\right)$. Under the conditions above, if we consider the liminf as $n \to \infty$ and take the worst case over I, we obtain the following local asymptotic minimax bound.

Theorem 2. *Assume the model* $\{Q_{n,\theta} : \theta \in \Theta\}$ *is locally asymptotically normal at* θ_0 *with scaling coefficient* $\sqrt{n}$ *and nonsingular* I_0*, that* $\theta_0 \in \text{int}(\Theta)$*, and that for all* $h \in \mathbb{R}^p$*, the*

moments of $Y_{n,h} = \frac{1}{\sqrt{n}} \sum_i \psi(\theta_{n,h}, X_i)$ *up to order* M *converge to the corresponding moments of* $\xi \sim N(0, \Omega)$*,*

$$E_{Q_{n,h}}\left[(v'Y_{n,h})^m\right] \to E\left[(v'\xi)^m\right] \quad \textit{for all } v \in \mathbb{R}^k \textit{ and all } m \in \{0, \ldots, M\}.$$

Then under Assumptions 1 and 2, for any sequence of decision rules δ_n*,*

$$\sup_I \liminf_{n\to\infty} \sup_{h\in I} \sup_{P \in \mathcal{P}^M_{n,h}} \left\{\mathbb{E}_P\left[L_n(\delta_n(X^n), \theta_{n,h})\right] - \lambda \,\mathrm{KL}(P \| Q_{n,h})\right\} \geq$$

$$\inf_\delta \sup_{h\in\mathbb{R}^p} \inf_{\beta\in\mathbb{R}^b} \lambda \cdot \log\left(E_{Q_h}\left[\ell^*\left(\delta(X,Y) - Kh\right)\exp\left(\beta' W_{M,h}\right)\right]\right),$$

where the supremum on the left-hand side ranges over finite subsets $I \subset \mathbb{R}^p$*,* $K = \frac{\partial}{\partial\theta'}\kappa(\theta_0) \in \mathbb{R}^{d\times p}$*,* Q_h *is given by*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} h \\ -\Psi h \end{pmatrix}, \begin{pmatrix} I_0^{-1} & -I_0^{-1}\Psi' \\ -\Psi I_0^{-1} & \Omega \end{pmatrix}\right), \tag{12}$$

and $W_{M,h}$ *collects the centered moments of* $Y_h = Y + \Psi h$ *up to order* M*,*

$$W_{M,h} = \left(\prod_{s=1}^k Y_{h,s}^{m_s} - E\left[\prod_{s=1}^k \xi_s^{m_s}\right] : m \in \mathbb{N}_0^k,\ 1 \leq \sum_{s=1}^k m_s \leq M\right). \tag{13}$$

Like the classical LAM theorem, Theorem 2 shows that the local asymptotic risk, now considering the constrained multiplier risk, is lower bounded by the risk in a Gaussian limit experiment. The limit experiment now involves two statistics: X, which plays the same role as in the standard LAM theorem and corresponds to the asymptotic analog of the maximum likelihood estimator, and Y, which is the limit of the scaled sample average $Y_{n,h}$ and captures the information in the moment conditions. The finite-sample constraint set $\mathcal{P}^M_{n,h}$, which requires that the first M moments of $Y_{n,h}$ be preserved, maps to the constraint $E_P[W_{M,h}] = 0$ in the limit experiment.

Example: Average Treatment Effects, Continued In the limit experiment (12), $X \in \mathbb{R}^2$ corresponds to the maximum likelihood estimator for (μ_0, μ_1) pooling data from both teams, while $Y \in \mathbb{R}^4$ captures the team-1 specific moment information. The matrices governing the

limit experiment are

$$I_0^{-1} = 2\,\mathrm{diag}(\sigma_0^2, \sigma_1^2), \quad \Psi = \begin{pmatrix} -\pi_1/2 & 0 \\ 0 & -\pi_1/2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Omega = \mathrm{diag}\left(\frac{\pi_1\sigma_0^2}{2}, \frac{\pi_1\sigma_1^2}{2}, \frac{\pi_1}{4}, \pi_1(1-\pi_1)\right).$$

Note that the third and forth rows of Ψ are zero, so the the corresponding elements of Y are uninformative on their own, but narrow the (asymptotic analog of the) ambiguity set. $\triangle$

When $M = 0$ there are no moment constraints, so the statistic Y plays no role: the adversary is free to distort its distribution. If ℓ is additionally symmetric around zero, the result reduces to the classical LAM theorem applied to the loss ℓ^*. Specifically, in this case $\ell^*(u) = \exp(\ell(u)/\lambda)$ is symmetric and (quasi-)convex, so the classical LAM theorem (Proposition 6) applies for each value of λ. More generally, the choice of M reflects what restrictions the researcher places on the forms of misspecification they consider. When the researcher is uncertain what value of M to impose, the $M \to \infty$ limit provides a natural benchmark.

Corollary 1. *If the assumptions of Theorem 2 hold for all $M \in \mathbb{N}$, then*

$$\inf_M \sup_I \liminf_{n\to\infty} \sup_{h\in I} \sup_{P\in\mathcal{P}_{n,h}^M} \{\mathbb{E}_P\left[L_n(\delta_n(X^n), \theta_{n,h})\right] - \lambda D_{KL}(P\|Q_{n,h})\} \geq$$

$$\inf_\delta \sup_h \lambda \cdot E_{Q_{Y,h}}\left[\log\left(E_{Q_{X|Y,h}}\left[\ell^*(\delta(X,Y) - Kh) \mid Y\right]\right)\right]$$

where Q_h is as in (12).

Taking $M \to \infty$ requires that misspecification not distort any moment of $Y_{n,h}$. Since the normal distribution is determined by its moments, this forces Y to remain normally distributed in the limit. The risk bound then simplifies because the KL divergence decomposes: the adversary can distort the conditional distribution of X given Y, but not the marginal of Y. The resulting bound has an intuitive form, with an inner conditional expectation of the exponentiated loss over $X \mid Y$, inside a logarithm, integrated over the marginal of Y.

Theorem 2 and Corollary 1 provide lower bounds on the risk achievable by any sequence of estimators under constrained multiplier preferences. In the next section, we characterize estimators which attain these bounds.

4 Optimal Decision Rules

To derive optimal estimators, we begin by exploiting the invariance structure of the limit experiment to show that we can limit attention to (asymptotically) equivariant decision rules. We then characterize optimal rules for important special cases and show that their finite-sample analogs, based on plugging the MLE and moments into the limit-experiment optimal rule, are asymptotically optimal.

4.1 The Hunt-Stein Theorem

The limit experiment (12) exhibits an important invariance structure. Consider the group $G = \mathbb{R}^p$ acting on the sample space by $g \circ (X, Y) = (X + g, Y - \Psi g)$, on the action space $\mathbb{R}^d$ by $g \circ a = a + Kg$, and on the parameter space by $g \circ h = h + g$. These transformations leave the loss unchanged: $\ell((g \circ a) - K(g \circ h)) = \ell(a - Kh)$ for all a, h, g. Following Lehmann and Casella (1998), we say that a decision rule δ is *equivariant* if $\delta(X+g, Y-\Psi g) = \delta(X, Y)+Kg$ for all $g \in \mathbb{R}^p$. Let $\mathcal{D}^E$ denote the class of equivariant decision rules in the limit experiment.

While the constrained multiplier objective is nonstandard, we extend the classical Hunt-Stein theorem to show that for minimax purposes, it is without loss to limit attention to equivariant decision rules.

Theorem 3. *Under Assumption 2, for any estimator δ in the limit problem there exists an equivariant estimator $\delta^E \in \mathcal{D}^E$ such that*

$$\sup_{h \in \mathbb{R}^p,\, P \in \mathcal{P}_{M,h}} E_P\left[\ell\left(\delta\left(X, Y\right) - Kh\right)\right] - \lambda \,\mathrm{KL}(P \| Q_h) \geq$$
$$\sup_{h \in \mathbb{R}^p,\, P \in \mathcal{P}_{M,h}} E_P\left[\ell\left(\delta^E\left(X, Y\right) - Kh\right)\right] - \lambda \,\mathrm{KL}(P \| Q_h).$$

Consequently, to derive a minimax decision rule it is without loss to restrict attention to the class of equivariant rules.

A useful consequence of equivariance is that the constrained multiplier risk does not depend on h. Specifically, for any $\delta^E \in \mathcal{D}^E$, transitivity of the group action implies

$$\sup_{P \in \mathcal{P}_{M,h}} E_P\left[\ell\left(\delta^E\left(X, Y\right) - Kh\right)\right] - \lambda \,\mathrm{KL}(P \| Q_h) = \sup_{P \in \mathcal{P}_{M,0}} E_P\left[\ell\left(\delta^E\left(X, Y\right)\right)\right] - \lambda \,\mathrm{KL}(P \| Q_0).$$

The minimax problem over equivariant rules thus reduces to

$$\inf_{\delta^E \in \mathcal{D}^E} \sup_{P \in \mathcal{P}_{M,0}} E_P \left[\ell \left(\delta^E (X, Y) \right) \right] - \lambda \, \mathrm{KL}(P \| Q_0).$$

Example: Average Treatment Effects, Continued In the two-team ATE example, the group $G = \mathbb{R}^2$ shifts (X, Y) by $(g, -\Psi g)$ and shifts the action by $Kg = (-1, 1)g$. Equivariance requires that the estimator respond to a location shift in the data (X, Y) by a parallel shift in the estimated treatment effect. The MLE $KX = (-1, 1)X$ is equivariant, as is the (limit experiment analog of the) efficient GMM estimator $-K(\Psi'\Omega^{-1}\Psi)^{-1}\Psi'\Omega^{-1}Y$. $\triangle$

4.2 Optimal Equivariant Rules

We next characterize optimal equivariant rules, treating the $M < \infty$ and $M = \infty$ cases in turn. By the duality of Proposition 5 and invariance, the minimax problem for equivariant rules with $M < \infty$ becomes

$$\inf_{\delta^E \in \mathcal{D}^E} \inf_{\beta \in \mathbb{R}^b} E_{Q_0} \left[\ell^* \left(\delta^E (X, Y) \right) \exp \left(\beta' W_{M,0} \right) \right]. \tag{14}$$

Since this is a joint infimum over (δ^E, β), the order of minimization does not matter. For each fixed β, however, the minimization over δ^E corresponds to finding the best equivariant decision rule under an exponentially tilted likelihood, since

$$\inf_{\delta^E \in \mathcal{D}^E} E_{Q_0} \left[\ell^* \left(\delta^E (X, Y) \right) \exp \left(\beta' W_{M,0} \right) \right] =$$

$$\inf_{\delta^E \in \mathcal{D}^E} E_{Q_0} \left[\ell^* \left(\delta^E (X, Y) \right) \frac{\exp \left(\beta' W_{M,0} \right)}{E_{Q_0} \left[\exp \left(\beta' W_{M,0} \right) \right]} \right] E_{Q_0} \left[\exp \left(\beta' W_{M,0} \right) \right].$$

Theorem 6.5 of Eaton (1989) proves that the best equivariant estimator in a location problem is the Bayes decision rule under the flat prior, from which we immediately obtain the form of the optimal rule for our problem.

Proposition 7. *Under Assumption 2, for each $\beta \in \mathbb{R}^b$ define*

$$\delta^*_\beta(X, Y) \in \arg\min_{a \in \mathcal{A}} \int \ell^*(a - Kh) \, \pi_\beta(h \mid X, Y) \, dh,$$

where $\pi_\beta(h \mid X, Y)$ *is the posterior density under a flat prior on* h*,*

$$\pi_\beta(h \mid X, Y) = \frac{q_0(X - h,\, Y + \Psi h) \cdot \exp\left(\beta' W_{M,h}\right)}{\int q_0(X - h',\, Y + \Psi h') \cdot \exp\left(\beta' W_{M,h'}\right) dh'},$$

and q_0 *denotes the density of* (X, Y) *under* Q_0*. Then:*

(a) δ^*_β *minimizes* $E_{Q_0}[\ell^*(\delta^E(X,Y)) \exp(\beta' W_{M,0})]$ *over* $\delta^E \in \mathcal{D}^E$*.*

(b) There exists β^* *minimizing*

$$\min_{\beta \in \mathbb{R}^b} E_{Q_0}\left[\ell^*\left(\delta^*_\beta(X,Y)\right) \exp\left(\beta' W_{M,0}\right)\right],$$

and moreover

(c) $\delta^*_{\beta^*}$ *is optimal in the limit experiment.*

The previous result provides the form of the optimal estimator for the case with $M < \infty$. In the case of $M = \infty$, the risk of an equivariant rule takes the form

$$R_\infty(\delta^E) = \lambda \cdot E_{Q_{Y,0}}\left[\log\left(E_{Q_{X|Y,0}}\left[\ell^*(\delta^E(X,Y)) \mid Y\right]\right)\right]. \tag{15}$$

We have not found a closed-form characterization of the optimal rule in the $M = \infty$ case under general loss. However, one can show that the risk (15) is convex in δ^E, so if we restrict to a linear class of rules $\delta_\Gamma(X,Y) = \Gamma'\phi(X,Y)$ for a finite-dimensional vector of basis functions $\phi(X,Y)$, the problem of finding the optimal rule in the class is likewise convex.

Proposition 8. *Under Assumption 2, let* $\phi(X,Y) \in \mathbb{R}^J$ *and let* $\Gamma \in \mathbb{R}^{d \times J}$*, so that* $\delta_\Gamma(X,Y) = \Gamma\phi(X,Y) \in \mathbb{R}^d$*.*

(a) For $M \in \mathbb{N}$*, the objective* $E_{Q_0}[\ell^*(\Gamma\phi(X,Y)) \exp(\beta' W_{M,0})]$ *is jointly convex in* (Γ, β)*.*

(b) For $M = \infty$*, the objective* $\lambda \cdot E_{Q_{Y,0}}[\log(E_{Q_{X|Y,0}}[\ell^*(\Gamma\phi(X,Y)) \mid Y])]$ *is convex in* Γ*.*

Using this convexity, it is straightforward to solve numerically for optimal rules in a given linear class provided one can tractably compute expectations under Q_0. For the important special case of squared error loss, however, we are able to go further exactly characterize the optimal decision rule for all M.

Proposition 9. *Under Assumption 2 with* $\ell(u) = \|u\|^2$*:*

(a) If $M = 0$ or $M = 1$, the optimal equivariant estimator is $\delta^*(X, Y) = KX$.

(b) If $M \geq 2$ (including $M = \infty$), the optimal equivariant estimator takes the form $\delta^*(X, Y) = K(X + C^* Z^I)$ for $Z^I = Y + \Psi X$, where

$$C^* \in \arg\min_C E_{Q_{Y,0}} \left[\log E_{Q_{X|Y,0}} \left[\exp\left(\frac{1}{\lambda} \left\| K(X + CZ^I) \right\|^2 \right) \mid Y \right] \right].$$

Moreover, the optimal risk is the same for all $M \geq 2$.

Part (b) implies that increasing the number of moment constraints M beyond two does not improve the optimal risk under squared error loss. Intuitively, we show that the best equivariant decision rule for $M = \infty$ is linear in (X, Y), but for such rules the worst-case distribution P^* is Gaussian, and a Gaussian distribution that matches the first two moments of Y automatically matches all higher moments as well.

Example: Average Treatment Effects, Continued In the two-team ATE example, part (a) says the optimal rule when $M \leq 1$ is KX, the MLE pooling data from both teams. Under $M \geq 2$, the researcher can exploit the team-1 moment conditions: the optimal rule linearly adjusts the MLE using $Z^I = Y + \Psi X$, which is the asymptotic analog to the moment conditions evaluated at the MLE. The optimal adjustment matrix C^* can be computed numerically. $\triangle$

4.3 Feasible, Asymptotically Optimal Rules

The results above characterize optimal decision rules in the limit experiment. In practice, however, the researcher observes only finite-sample data. We now show that plug-in finite-sample analogs of limit-experiment rules converge in both distribution and (under integrability conditions) risk.

For a continuous, equivariant decision rule $\delta^c(X, Y; \Sigma)$ in the limit experiment, define a plug-in finite-sample analog as

$$\delta_n^c = \kappa(\hat{\theta}_n^{MLE}) + \frac{1}{\sqrt{n}} \delta^c \left(0,\ \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(\hat{\theta}_n^{MLE}, X_i);\ \hat{\Sigma}_n \right),$$

where $\hat{\theta}_n^{MLE}$ is the maximum likelihood estimator and $\hat{\Sigma}_n$ is a consistent estimator of the limit covariance matrix.

Assumption 3. *For ψ_j the j-th component of ψ, there exist $\eta > 0$ and $\tau_{j,k,l}(X)$ such that*

$$\sup_{\theta \in \mathcal{N}} \left| \frac{\partial^2}{\partial \theta_k \partial \theta_l} \psi_j(\theta, X) \right| \leq \tau_{j,k,l}(X)$$

for all j, k, l, where $\mathcal{N}$ is a neighborhood of θ_0 and $E_{Q_{n,\theta_0}}[\tau_{j,k,l}(X_i)^{1+\eta}] < C < \infty$. Moreover, for every $h \in \mathbb{R}^p$,

$$\left(\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_{n,h}),\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\hat{\theta}_n^{MLE}, X_i),\ \hat{\Sigma}_n \right) \overset{Q_{n,h}}{\underset{d}{\to}} (X - h,\ Y + \Psi X,\ \Sigma)\,.$$

Proposition 10. *Under Assumptions 1–3 and the conditions of Theorem 2, let $\delta^c(X, Y; \Sigma)$ be a continuous equivariant decision rule in the limit experiment. Then for any $h \in \mathbb{R}^p$,*

$$\sqrt{n}\left(\delta_n^c - \kappa(\theta_{n,h})\right) \overset{Q_{n,h}}{\underset{d}{\to}} \delta^c(X, Y; \Sigma) - Kh.$$

To obtain convergence of the (dual) constrained multiplier objective, we strengthen convergence in distribution to convergence of moments.

Assumption 4. *For continuous equivariant decision rule δ^c in the limit experiment, let*

$$\beta^{*,c} \in \arg\min_{\beta \in \mathbb{R}^b} E_{Q_0}\left[\ell^*\left(\delta^c(X, Y; \Sigma)\right) \exp\left(\beta' W_{M,0}\right)\right].$$

For each finite $h \in \mathbb{R}^p$,

$$E_{Q_{n,h}}\left[\ell^*\left(\sqrt{n}\left(\delta_n^c - \kappa(\theta_{n,h})\right)\right) \exp\left(\beta^{*,c\prime} W_{M,n,h}\right)\right] \to E_{Q_h}\left[\ell^*\left(\delta^c(X, Y; \Sigma) - Kh\right) \exp\left(\beta^{*,c\prime} W_{M,h}\right)\right].$$

Corollary 2. *Under Assumptions 1–4 and the conditions of Theorem 2, for continuous, equivariant decision rule δ^c in the limit experiment and δ_n^c its plug-in analog,*

$$\lim_{n\to\infty} \sup_{h \in I} \inf_{\beta} \lambda \cdot \log\left(E_{Q_{n,h}}\left[\ell^*\left(\sqrt{n}(\delta_n^c - \kappa(\theta_{n,h}))\right) \exp(\beta' W_{M,n,h})\right]\right) =$$

$$\sup_{h \in I} \inf_{\beta} \lambda \cdot \log\left(E_{Q_h}\left[\ell^*\left(\delta^c(X, Y; \Sigma) - Kh\right) \exp(\beta' W_{M,h})\right]\right)$$

for any finite $I \subset \mathbb{R}^p$. In particular, if δ^c is an optimal rule from Section 4.2, then the

plug-in estimator δ_n^c is asymptotically optimal:

$$\sup_I \liminf_{n\to\infty} \sup_{h\in I} \sup_{P\in\mathcal{P}_{n,h}^M} \{\mathbb{E}_P\left[L_n(\delta_n^c(X^n),\theta_{n,h})\right] - \lambda D_{KL}(P\|Q_{n,h})\} =$$

$$\inf_\delta \sup_h \inf_\beta \lambda \cdot \log\left(E_{Q_h}\left[\ell^*\left(\delta(X,Y) - Kh\right)\exp(\beta' W_{M,h})\right]\right).$$

5 Adaptive Decision Rules

While the results in Section 4 characterize optimal decision rules under constrained multiplier risk, the resulting rules depend on the misspecification-aversion parameter λ. While dependence of optimal rules on λ is natural from a theoretical perspective (after all, a researcher entirely unconcerned with misspecification already knows the MLE is optimal), from a practical perspective it introduces a free parameter that a researcher interested in applying our methods must choose. In this section, we follow Armstrong et al. (2025) and examine, for a special case of our setting, whether there exist simple estimators that perform reasonably for many different values of λ.

In particular, we consider the special case of squared error loss $\ell = \|u\|^2$ where both the parameter θ and the moment condition ψ are scalar, where we normalize $I_0 = 1$ and $\Psi = -1$, so the limit problem is fully described by $\Omega = \mathrm{Var}(Y)$.[9] We take $M = \infty$, corresponding to the case where the researcher thinks misspecification does not affect the moments at all. Thus, the sole remaining decision for the researcher is the (unavoidable) choice of which moment function ψ they think remains valid under misspecification.

To search for procedures which perform well across different values of λ, let $R_\infty^\lambda(\delta)$ denote the constrained multiplier risk of δ under misspecification-aversion parameter λ,

$$R_\infty^\lambda(\delta) = \sup_{h\in\mathbb{R}^p,\, P\in\mathcal{P}_{\infty,h}} E_P\left[\ell\left(\delta\left(X,Y\right) - Kh\right)\right] - \lambda\,\mathrm{KL}(P\|Q_h).$$

[9] As in Armstrong et al. (2025), one could more broadly interpret this setting as the limit problem when we are interested in a scalar target parameter and would like to combine an efficient but potentially misspecified estimator (represented by X) and less efficient but more robust estimator (represented by Y). The distinction between our analysis in this section and that of Armstrong et al. (2025) then stems from their focus on a version of constraint preferences, vs. ours on constrained multiplier preferences.

Following Armstrong et al. (2025) we consider the *adaptive regret* criterion

$$A_\infty(\delta) = \sup_{\lambda \in \Lambda} \frac{R^\lambda_\infty(\delta)}{\min_{\tilde{\delta}} R^\lambda_\infty(\tilde{\delta})},$$

which compares the performance of the rule δ, across values of λ, to the performance of the family of λ-by-λ optimal rules. By construction $A_\infty(\delta) \geq 1$, where if the adaptive regret is close to this lower bound it tells us that the rule δ is "nearly" optimal in a proportional sense uniformly across λ values, and thus that if we opt to use δ rather than taking a stand on the "correct" λ and using the resulting rule, the price of doing so (measured as the proportional increase in constrained multiplier risk) is not very high.

By Proposition 9, we know that to solve for the λ-specific minimized risk $\min_{\tilde{\delta}} R^\lambda_\infty(\tilde{\delta})$, it suffices to consider linear equivariant rules, greatly facilitating computation. Analogously, in the broader search for adaptive rules, we limit attention to equivariant rules $\delta^E \in \mathcal{D}^E$. Let us normalize $I_0 = 1$ and $\Psi = -1$, and again define $Z^I = Y + \Psi X$ as the limit experiment analog of the sample average moments evaluated at the MLE, or (equivalently) as the difference between the GMM and ML estimators. One can show that Z^I is a maximal invariant in the limit experiment, and thus that any equivariant decision rule can be written as

$$\delta^E(X, Y) = X + \gamma(Z^I).$$

Thus, the problem of finding an equivariant estimator with a small adaptive risk is equivalent to picking a "good" γ.

We consider two simple parameterizations of $\gamma(\cdot)$ with a single tuning parameter, both of which Armstrong et al. (2025) find perform well according to their criterion. The first the soft-thresholding estimator with

$$\gamma_{ST,\tau}(Z^I) = \max\left\{|Z^I| - \tau, 0\right\} \operatorname{sgn}(Z^I),$$

which is closely related to the LASSO estimator (Tibshirani, 1996). The second is the adaptive empirical risk minimization (ERM) estimator (Magnus, 2002; de Chaisemartin and D'Haultfœuille, 2020) with

$$\gamma_{ERM,\tau}(Z^I) = \frac{(Z^I)^2}{(Z^I)^2 + \tau} Z^I.$$

In each class, we choose the tuning parameter τ to minimize the adaptive risk A_∞. To benchmark the performance of these simple parametric classes, we compare their adaptive

risk to that of a flexible specification of γ. In particular, motivated by Proposition 8, we parameterize γ as a cubic spline and solve numerically for the optimal parameters via convex optimization.[10]

Figure 1(a) shows the resulting constrained multiplier risk functions for the λ-by-λ optimal estimator, the adaptive cubic spline (pchip) estimator, and the two simple estimators, all for the case of $\Omega = 2$. As expected, for small λ the optimal risk is close to that of GMM (i.e. $\Omega = 2$) while for large λ is is close to that of MLE (i.e. $I_0 = 1$). In between, we see that the cubic spline estimator has better performance than the simple estimators over an intermediate range of λ values, but performs quite similarly for large and small λ. Figure 1(b) illustrates adaptive performance more directly, plotting the risk ratio $\frac{R^\lambda_\infty(\delta)}{\min_{\tilde{\delta}} R^\lambda_\infty(\tilde{\delta})}$ as a function of λ. Here we see that, consistent with the findings of Armstrong et al. (2025) for their optimality criterion, the simple soft-thresholding and ERM estimators perform nearly as well as the more complicated cubic spline procedure, with an adaptive risk close to 1.5 for all procedures considered. Whether a 50% increase in constrained multiplier risk is an acceptable tradeoff for eliminating dependence on λ seems to depend on one's priorities, but we find it encouraging that near-optimal adaptation is possible using simple combinations of the ML and GMM estimators.

References

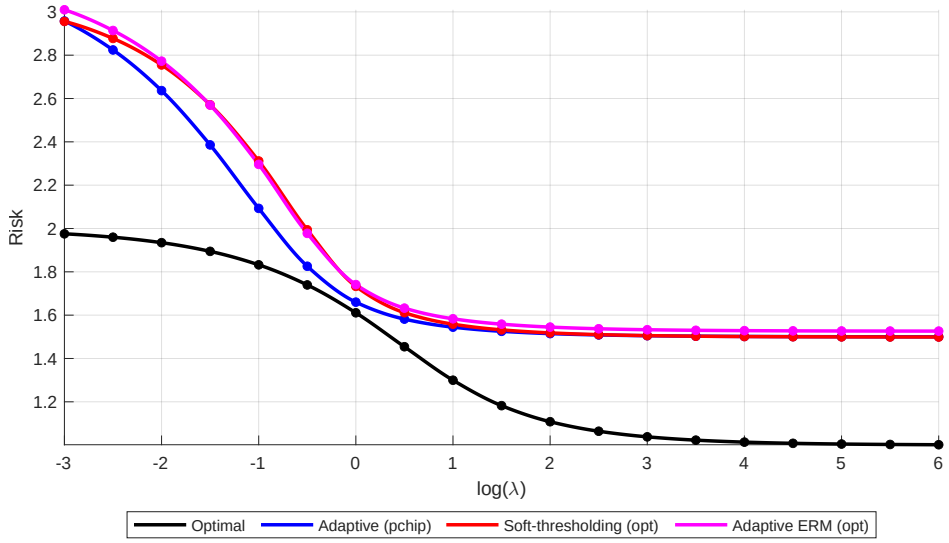
Aliprantis, C. D., and K. C. Border (2006): *Infinite dimensional analysis: a hitchhiker's guide*: Springer.

Andrews, I., J. Chen, and O. Tecchio (2025): "The purpose of an estimator is what it does: Misspecification, estimands, and over-identification," *arXiv preprint arXiv:2508.13076*.

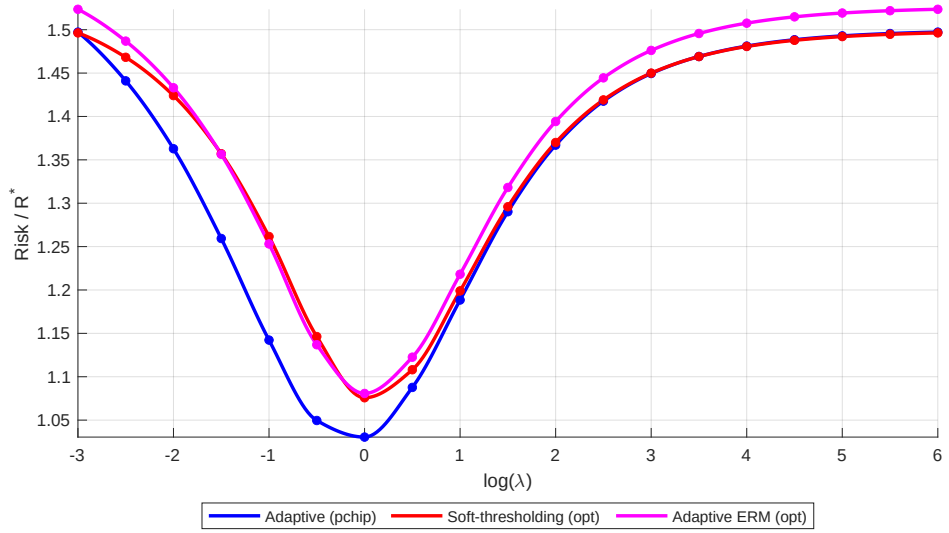
Armstrong, T. B., P. Kline, and L. Sun (2025): "Adapting to Misspecification," *Econometrica*, 93, 1981–2005.

Armstrong, T. B., and M. Kolesár (2021): "Sensitivity Analysis Using Approximate Moment Condition Models," *Quantitative Economics*, 12, 1–39.

[10]Specifically, the maximization which defines A_∞ preserves the convexity established by Proposition 8, and we approximate the maximum over λ by maximization over the finite, evenly spaced grid of $\log(\lambda)$ values from [-3,6].



(a) Risk of Adaptive Estimators



(b) Risk Ratio against Pointwise Optimum

Figure 1: Adaptively Optimal, Soft-thresholding and ERM Estimators ($\Omega = 2$)

Bonhomme, S., and M. Weidner (2022): "Minimizing Sensitivity to Model Misspecification," *Quantitative Economics*, 13, 641–685.

Breza, E., A. G. Chandrasekhar, and D. Viviano (2025): "Generalizability with ignorance in mind: learning what we do (not) know for archetypes discovery."

Cerreia-Vioglio, S., L. P. Hansen, F. Maccheroni, and M. Marinacci (2025): "Making Decisions under Model Misspecification," *Review of Economic Studies*, Forthcoming.

de Chaisemartin, C., and X. D'Haultfœuille (2020): "Empirical MSE Minimization to Estimate a Scalar Parameter."

Christensen, T., and B. Connault (2023): "Counterfactual Sensitivity and Robustness," *Econometrica*, 91, 263–298.

Cobasz, S., R. Miculescu, and A. Nicolae (2019): *Lipschitz Functions*: Springer.

Eaton, M. L. (1989): *Group Invariance Applications in Statistics*, Hayward, CA: Institute of Mathematical Statistics and American Statistical Association.

Ghirardato, P., and M. Marinacci (2002): "Ambiguity Made Precise: A Comparative Foundation," *Journal of Economic Theory*, 102, 251–289.

Gilboa, I., and D. Schmeidler (1989): "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 18, 141–153.

Hájek, J. (1972): "Local Asymptotic Minimax and Admissibility in Estimation," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* Volume 1: University of California Press, 175–194.

Hansen, L. P., and T. J. Sargent (2001): "Robust Control and Model Uncertainty," *American Economic Review*, 91, 60–66.

——— (2008): *Robustness*: Princeton University Press.

Kitamura, Y. (2009): *Empirical Likelihood Methods in Econometrics: Theory and Practice*, Chap. 7: Cambridge University Press.

Lanzani, G. (2025): "SUPPLEMENT TO "DYNAMIC CONCERN FOR MISSPECIFICATION"," *Econometrica Supple*.

Lehmann, E. L., and G. Casella (1998): *Theory of Point Estimation*, New York: Springer, 2nd edition.

Maccheroni, F., M. Marinacci, and A. Rustichini (2006): "Ambiguity aversion, robustness, and the variational representation of preferences," *Econometrica*, 74, 1447–1498.

Magnus, J. (2002): "Estimation of the mean of a univariate normal distribution with known variance," *The Econometrics Journal*, 5, 225–236.

Newey, W. K., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255.

Stoye, J. (2012): "New Perspectives on Statistical Decisions Under Ambiguity," *Annual Review of Economics*, 4, 575–595.

Strzalecki, T. (2011): "Axiomatic foundations of multiplier preferences," *Econometrica*, 79, 47–73.

Tibshirani, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

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

Zălinescu, C. (2002): *Convex analysis in general vector spaces*: World Scientific.

A Axioms for Misspecification-Averse Preferences

We begin by presenting a series of axioms and corresponding representation theorems adapted to our loss function setting, which characterize the preferences (1) and (2). While the results in this appendix largely follow from arguments in Maccheroni et al. (2006) and Cerreia-Vioglio et al. (2025), we hope that our framing of the choice problem in terms of loss functions may be helpful for readers with a background in econometrics and statistics.

A.1 Basic Axioms

We begin by imposing a set of regularity conditions on $\succsim_\Theta$ and $\succsim_\theta$. Recall that these are binary relations on $\mathcal{L}$, which is the set of bounded, Borel measurable functions $L : \Theta \times \mathcal{X} \to \mathbb{R}$. To state the first axiom, let $L_\theta = L(\theta, \cdot) : \mathcal{X} \to \mathbb{R}$ be the loss function L conditional on θ. For $L \in \mathcal{L}$ and $\theta \in \Theta$, we will abuse notation and let $L_\theta \in \mathcal{L}$ refer to the loss function satisfying: $L_\theta(\theta', x) = L(\theta, x)$ for all $\theta' \in \Theta$ and $x \in \mathcal{X}$.

Axiom 6 (θ-Relevance)**.** *For all $\theta \in \Theta$ and all $L, L' \in \mathcal{L}$,*

$$L \succsim_\theta L' \iff L_\theta \succsim_\theta L'_\theta$$

θ-Relevance ensures that $\succsim_\theta$ only cares about loss functions conditional on θ.

Axiom 7 (Nontrivial Weak Order)**.** *The relation $\succsim_\Theta$ is complete, transitive, and nontrivial. For each $\theta \in \Theta$, the relation $\succsim_\theta$ is complete, transitive, and nontrivial.*

Completeness requires that the researcher be able to rank any two loss functions. Transitivity is a standard coherence requirement. Nontriviality rules out the uninteresting case where the researcher is indifferent between all loss functions.

Axiom 8 (Monotonicity)**.** *If $L(\theta', x) \leq L'(\theta', x)$ for all $(\theta', x) \in \Theta \times \mathcal{X}$, then $L \succsim_\Theta L'$. For each $\theta \in \Theta$, if $L(\theta, x) \leq L'(\theta, x)$ for all $x \in \mathcal{X}$, then $L \succsim_\theta L'$.*

Monotonicity requires that the researcher prefer lower losses: if L yields weakly lower loss than L' in every state, then the researcher must weakly prefer L.[11] This is a weak dominance condition satisfied by any sensible loss-based criterion.

Next, we state several continuity axioms. Such axioms are standard to derive utility representations, but we view them as primarily technical conditions. Our first continuity axiom is imposed on $\succsim_\Theta$.

Axiom 9 (Θ-Mixture Continuity)**.** *For $L, L', L'' \in \mathcal{L}$, the sets $\{\alpha \in [0,1] : \alpha L + (1-\alpha)L' \succsim_\Theta L''\}$ and $\{\alpha \in [0,1] : L'' \succsim_\Theta \alpha L + (1-\alpha)L'\}$ are closed.*

Mixture Continuity ensures that small perturbations to mixtures of fixed loss functions do not lead to discontinuous jumps in the preference ranking. Here and throughout, the mixture $\alpha L + (1-\alpha)L'$ is defined pointwise: $[\alpha L + (1-\alpha)L'](\theta, x) = \alpha L(\theta, x) + (1-\alpha)L'(\theta, x)$. Such

[11] Note that Monotonicity implies: for each $\succsim \in \{\succsim_\Theta, \{\succsim_\theta\}_{\theta \in \Theta}\}$, $\succsim$ restricted to $\mathbb{R}$ satisfies: $r \leq r'$ if and only if $r \succsim r'$. In particular, this implies Maccheroni et al. (2006)'s Axiom A.7 (Unboundedness).

mixtures arise naturally when the researcher randomizes between decision rules: if they use δ with probability α and δ' otherwise, independent of the data, then the induced loss is $\alpha L_\delta + (1-\alpha)L_{\delta'}$.

Our second notion of continuity, imposed on $\{\succsim_\theta : \theta \in \Theta\}$, is stronger and requires a topology on the space of loss functions. We endow $\mathcal{L}$ with the sup-norm topology.

Axiom 10 (θ-Continuity)**.** *For each $L \in \mathcal{L}$, the sets*

$$\{L' \in \mathcal{L} : L' \succsim_\theta L\} \quad \textit{and} \quad \{L' \in \mathcal{L} : L \succsim_\theta L'\}$$

are closed.

θ-Continuity ensures that small perturbations (in the sense of sup-norm) of loss functions do not imply discontinuous jumps in the preference ranking. Note that, since $\alpha_n \to \alpha$ implies $\alpha_n L + (1-\alpha_n)L' \to \alpha L + (1-\alpha)L'$, θ-Continuity implies θ-Mixture Continuity.

Axiom 10' (θ-Mixture Continuity)**.** *For each $\theta \in \Theta$ and $L, L', L'' \in \mathcal{L}$, the sets $\{\alpha \in [0,1] : \alpha L + (1-\alpha)L' \succsim_\theta L''\}$ and $\{\alpha \in [0,1] : L'' \succsim_\theta \alpha L + (1-\alpha)L'\}$ are closed.*

A.2 Axiomatization of Variational Preferences

We next present axioms that imply a particular functional form for the conditional preferences $\succsim_\theta$. As we discuss in the main text, this class of *variational preferences* nests a number of important cases previously discussed in the literature on estimation and decision-making with misspecification concerns.

The first axiom restricts how $\succsim_\theta$ responds to mixing with constant losses.

Axiom 11 (Weak Certainty Independence)**.** *For all $\theta \in \Theta$, $L, L' \in \mathcal{L}$, $r, r' \in \mathbb{R}$, and $\alpha \in (0,1)$,*

$$\alpha L + (1-\alpha)r \succsim_\theta \alpha L' + (1-\alpha)r \Rightarrow \alpha L + (1-\alpha)r' \succsim_\theta \alpha L' + (1-\alpha)r'$$

Weak Certainty Independence requires that preferences between loss functions mixed with constant acts not depend on the value of the constant act.

Axiom 12 (Uncertainty Aversion)**.** *For all $\theta \in \Theta$, $L, L' \in \mathcal{L}$, and $\alpha \in (0,1)$,*

$$L \sim_\theta L' \Rightarrow \alpha L + (1-\alpha)L' \succsim_\theta L$$

Uncertainty Aversion requires that the researcher have a weak preference for hedging: if they are indifferent between two loss functions, they must weakly prefer a mixture of the two. This captures the intuition that diversification (e.g. via randomization) is valuable when facing uncertainty about the data generating process.

The remaining axioms imply certain properties of the cost function. First, we require that $\succsim_\theta$ always finds Q_θ at least as plausible as any other DGP.

Axiom 13. *For all $\theta \in \Theta$, $\succsim_\theta$ is more ambiguity averse than $\succsim^{SEU}_{\theta,Q_\theta}$.*

Let $\Delta^F(\mathcal{X})$ be the set of finitely additive Borel probability measures on $\mathcal{X}$, endowed with the weak*-topology. Recall that $\Delta(\mathcal{X}) \subseteq \Delta^F(\mathcal{X})$ is the set of countably additive Borel probability measures on $\mathcal{X}$. For a set $C \subseteq \Delta^F(\mathcal{X})$, let $\bar{C}$ denote its weak*-closure. Note that, under the axioms we have imposed so far, for each $\theta \in \Theta$ and $L \in \mathcal{L}$, there exists a unique $CE_\theta(L) \in \mathbb{R}$ such that $CE_\theta(L) \sim_\theta L$.[12] Let $\mathcal{L}_0$ denote the set of simple, measurable loss functions. We will show that the following axiom ensures that the variational representation may be written only in the language of countably additive probabilities.

Axiom 14. For all $\theta \in \Theta$ and $t \geq 0$,

$$\left\{P \in \Delta^F(\mathcal{X}) : \sup_{L \in \mathcal{L}_0} \left(\int L_\theta \, dP - CE_\theta(L) \right) \leq t \right\} = \overline{\left\{P \in \Delta(\mathcal{X}) : \sup_{L \in \mathcal{L}_0} \left(\int L_\theta \, dP - CE_\theta(L) \right) \leq t \right\}}$$

Finally, the following axiom ensures that the cost function is weakly lower semicontinuous.

Axiom 15. *For all $\theta \in \Theta$ and $t \geq 0$, the set*

$$\left\{P \in \Delta(\mathcal{X}) : \sup_{L \in \mathcal{L}_0} \left(\int L_\theta \, dP - CE_\theta(L) \right) \leq t \right\}$$

is weakly closed.

Remark. We may also endow $\Delta(\mathcal{X})$ with the topology of weak convergence. Under the axioms we have imposed so far, note that Axioms 14 and 15 are implied by Monotone Continuity, since by Theorem 13 of Maccheroni et al. (2006), Monotone Continuity implies that the LHS set of Axiom 14 is a weakly compact (and hence weakly closed) subset of $\Delta(\mathcal{X})$.

[12] The argument is exactly analogous to the proof of Step 1 of Theorem 5(i).

Definition 2. *The preference $\succsim_\theta$ has a* variational representation *on $\mathcal{L}$ if there exists a convex, weak* lower-semicontinuous function $c_\theta : \Delta^F(\mathcal{X}) \to [0, \infty]$ with $\inf_{P \in \Delta^F(\mathcal{X})} c_\theta(P) = 0$ such that*

$$V_\theta(L) = \max_{P \in \Delta^F(\mathcal{X})} \left\{ \int L_\theta(x) dP(x) - c_\theta(P) \right\} \tag{16}$$

represents $\succsim_\theta$ on $\mathcal{L}$, in the sense that $L \succsim_\theta L' \iff V_\theta(L) \leq V_\theta(L')$ for all $L, L' \in \mathcal{L}$.

Note that the max in Equation (16) is attained because $\Delta^F(\mathcal{X})$ is weak*-compact and the map

$$P \mapsto \int L_\theta \, dP - c_\theta(P)$$

is the sum of a real-valued, weak* continuous function $P \mapsto \int L_\theta \, dP$ (since L_θ is bounded and measurable) and a weak* upper-semicontinuous function $P \mapsto -c_\theta(P)$.

Our main result in this section establishes that, under the axioms we have imposed thus far, the conditional preferences $\succsim_\theta$ have a variational representation whose cost function c_θ satisfies certain appealing properties. Many of the representation results in the main text refine this one by characterizing particular functional forms of the cost function c_θ.

Theorem 4. *(i) The preference $\succsim_\theta$ satisfies Axioms 6–12 if and only if it has a variational representation on $\mathcal{L}$.*

(ii) If $\succsim_\theta$ has a variational representation on $\mathcal{L}$, it is unique and given by:

$$c_\theta(P) = \sup_{L \in \mathcal{L}_0} \left(\int L \, dP - CE_\theta(L) \right)$$

(iii) Suppose $\succsim_\theta$ satisfies Axioms 6–12. It additionally satisfies Axiom 13 if and only if $c_\theta(Q_\theta) = 0$.

(iv) Suppose $\succsim_\theta$ satisfies Axioms 6–12. It additionally satisfies Axiom 14 if and only if

$$\{P \in \Delta^F(\mathcal{X}) : c_\theta(P) \leq t\} = \overline{\{P \in \Delta(\mathcal{X}) : c_\theta(P) \leq t\}} \quad \forall t \geq 0$$

In this case, for V_θ as defined in Equation (16),

$$V_\theta(L) = \sup_{P \in \Delta(\mathcal{X})} \left\{ \int L_\theta(x) \, dP(x) - c_\theta(P) \right\} \quad \forall L \in \mathcal{L}$$

In particular for $L = 0$,

$$\inf_{P \in \Delta(\mathcal{X})} c_\theta(P) = 0$$

(v) Suppose $\succsim_\theta$ *satisfies Axioms 6–12. It additionally satisfies Axiom 15 if and only if the restriction of* c_θ *to* $\Delta(\mathcal{X})$ *is weakly lower-semicontinuous.*

Before proving Theorem 4, we prove two useful lemmas which we use throughout Appendices A and B. These two lemmas will allow us to 1) equivalently work with utility representations rather than risk representations; and 2) extend representation theorems for preferences restricted to $\mathcal{L}_0$ to preferences over all of $\mathcal{L}$. Hence, these lemmas address two features of our choice environment that are nonstandard in the microeconomic decision theory literature: our primitives are preferences over bounded measurable loss functions rather than simple measurable Anscombe–Aumann acts.

First, we observe that working with preferences over loss functions is equivalent to working with preferences over negative loss functions, or *utility acts*. Recall that a function V_θ *represents* $\succsim_\theta$ *on* $\mathcal{L}$ if and only if

$$L \succsim_\theta L' \iff V_\theta(L) \leq V_\theta(L') \quad \forall L, L' \in \mathcal{L}$$

For each loss function $L \in \mathcal{L}$, define its induced *utility act* to be $f_L = -L$. Let $\mathcal{F} = \{f_L : L \in \mathcal{L}\}$ be the set of utility acts induced by some (bounded, Borel measurable) loss function, and note that $\mathcal{F}$ is precisely the set of bounded, Borel measurable functions $f : \Theta \times \mathcal{X} \to \mathbb{R}$. Finally, for each $\succsim_\theta$, define the binary relation $\succsim_\theta^{\mathcal{F}}$ on $\mathcal{F}$ as:

$$f_L \succsim_\theta^{\mathcal{F}} f_{L'} \iff L \succsim_\theta L'$$

For a binary relation $\succsim'_\theta$ on $\mathcal{F}$ and a function $U_\theta : \mathcal{F} \to \mathbb{R}$, say that U_θ *represents* $\succsim'_\theta$ *on* $\mathcal{F}$ if and only if

$$f \succsim'_\theta g \iff U_\theta(f) \geq U_\theta(g) \quad \forall f, g \in \mathcal{F}$$

Lemma 1. *V_θ represents $\succsim_\theta$ on $\mathcal{L}$ if and only if $U_\theta = -V_\theta(-\cdot)$ represents $\succsim_\theta^{\mathcal{F}}$ on $\mathcal{F}$.*

Proof of Lemma 1. The following statements (i)-(v) are equivalent:

(i) V_θ represents $\succsim_\theta$ on $\mathcal{L}$.

(ii) $L \succsim_\theta L' \iff V_\theta(L) \leq V_\theta(L') \quad \forall L, L' \in \mathcal{L}$.

(iii) $-L \succsim_\theta^{\mathcal{F}} -L' \iff -U_\theta(-L) \leq -U_\theta(-L') \quad \forall L, L' \in \mathcal{L}$.

(iv) $f \succsim_\theta^{\mathcal{F}} g \iff U_\theta(f) \geq U_\theta(g) \quad \forall f, g \in \mathcal{F}$.

(v) U_θ represents $\succsim_\theta^{\mathcal{F}}$ on $\mathcal{F}$.

□

In particular, Lemma 1 implies: V_θ as defined in Equation (16) represents $\succsim_\theta$ on $\mathcal{L}$ if and only if $U_\theta : \mathcal{F} \to \mathbb{R}$ defined as:

$$U_\theta(f) = \min_{P \in \Delta^F(\mathcal{X})} \left\{ \int f_\theta(x) \, dP(x) + c_\theta(P) \right\} \tag{17}$$

represents $\succsim_\theta^{\mathcal{F}}$ on $\mathcal{F}$. Note that for each axiom we have defined on $\succsim_\theta$, there exists an corresponding, equivalent axiom on $\succsim_\theta^{\mathcal{F}}$. Moving forward, we abuse notation and use $\succsim_\theta$ to also refer to its induced preference on utility acts $\succsim_\theta^{\mathcal{F}}$.

Second, we establish that to obtain a variational representation U_θ (as defined in Equation 17) of $\succsim_\theta$ on $\mathcal{F}$, it suffices to obtain a variational representation U_θ of $\succsim_\theta$ on the domain $\mathcal{F}_0(\mathcal{X}) \subseteq \mathcal{F}$ of *simple* (finitely-valued) Borel measurable functions f such that $f(\cdot, x)$ is constant in θ. We abuse notation and identify this set with the set of simple, Borel measurable functions $f : \mathcal{X} \to \mathbb{R}$.

Lemma 2. *Assume that $\succsim_\theta$ satisfies the basic axioms (θ-Relevance, Nontrivial Weak Order, Monotonicity, and θ-Continuity), and $\succsim_\theta$ has a variational representation on $\mathcal{F}_0(\mathcal{X})$: there exists a convex, weak* lower-semicontinuous function $c_\theta : \Delta^F(\mathcal{X}) \to [0, \infty]$ with $\inf_{P \in \Delta^F(\mathcal{X})} c_\theta(P) = 0$ such that*

$$f \succsim_\theta g \iff U_\theta(f) \geq U_\theta(g) \quad \forall f, g \in \mathcal{F}_0(\mathcal{X})$$

where U_θ is defined as in Equation (17). Then, the above equivalence holds for all $f, g \in \mathcal{F}$, and hence U_θ represents $\succsim_\theta$ on $\mathcal{F}$.

Proof of Lemma 2. Step 1: U_θ is continuous on $\mathcal{F}$. Define $U_\theta : \mathcal{F} \to \mathbb{R}$ as in Equation (17). Fix any $f, g \in \mathcal{F}$ and without loss of generality, suppose $U_\theta(f) \geq U_\theta(g)$. Choose $P \in \Delta^F(\mathcal{X})$ such that

$$U_\theta(g) = \int g_\theta \, dP + c_\theta(P)$$

Hence,

$$\begin{aligned} U_\theta(f) - U_\theta(g) &\leq \int f_\theta \, dP + c_\theta(P) - \int g_\theta \, dP - c_\theta(P) \\ &= \int (f_\theta - g_\theta) \, dP \leq \int |f_\theta - g_\theta| \, dP \leq \|f_\theta - g_\theta\|_\infty \end{aligned}$$

where the last two inequalities follow from monotonicity of the integral against finitely additive measures. Hence, U_θ is 1-Lipschitz and hence continuous on $\mathcal{F}$.

Step 2: U_θ represents $\succsim_\theta$ on $\mathcal{F}$. Fix any $f, g \in \mathcal{F}$.

First, suppose that $f \succsim_\theta g$. By θ-Relevance, $f_\theta \succsim_\theta g_\theta$. For each $k \geq 1$ and $j \geq 1$, define

$$f_{\theta,k}(x) = \frac{\lfloor k f_\theta(x) \rfloor + 1}{k} \quad \text{and} \quad g_{\theta,j}(x) = \frac{\lfloor j g_\theta(x) \rfloor}{j}$$

Since f_θ, g_θ are bounded, each $f_{\theta,k}, g_{\theta,j}$ is simple. Also note that $k f_\theta(x) \leq \lfloor k f_\theta(x) \rfloor + 1$ and $\lfloor j g_\theta(x) \rfloor \leq j g_\theta(x)$ imply $f_\theta \leq f_{\theta,k}$ and $g_\theta \geq g_{\theta,j}$. Finally, note that $\|f_{\theta,k} - f_\theta\|_\infty \leq k^{-1} \to 0$ and $\|g_{\theta,j} - g_\theta\|_\infty \leq j^{-1} \to 0$.

For each fixed $j \geq 1$, Monotonicity implies:

$$f_{\theta,k} \succsim_\theta f_\theta \succsim_\theta g_\theta \succsim_\theta g_{\theta,j} \quad \forall k \geq 1 \implies U_\theta(f_{\theta,k}) \geq U_\theta(g_{\theta,j}) \quad \forall k \geq 1$$

and hence

$$U_\theta(f_\theta) = \lim_{k \to \infty} U_\theta(f_{\theta,k}) \geq U_\theta(g_{\theta,j})$$

Finally, taking $j \to \infty$ yields $U_\theta(f_\theta) \geq U_\theta(g_\theta)$. By relabeling f_θ and g_θ, we also see that if $f_\theta \sim_\theta g_\theta$, then $U_\theta(f_\theta) = U_\theta(g_\theta)$.

Finally, suppose that $f \succ_\theta g$. By θ-Relevance, $f_\theta \succ_\theta g_\theta$. By θ-continuity, there exists $\epsilon > 0$ small enough such that $f_\theta \succ_\theta g_\theta + \epsilon$. Define $(f_{\theta,k})_k$ and $(g_{\theta,j})_j$ as before. For each fixed $j \geq 1$, Monotonicity implies:

$$f_{\theta,k} \succsim_\theta f_\theta \succ_\theta g_\theta + \epsilon \succsim_\theta g_{\theta,j} + \epsilon \quad \forall k \geq 1 \implies U_\theta(f_{\theta,k}) > U_\theta(g_{\theta,j} + \epsilon) = U_\theta(g_{\theta,j}) + \epsilon \quad \forall k \geq 1$$

where the last equality uses the fact that U_θ is translation invariant. Hence, taking $k \to \infty$ yields:

$$U_\theta(f_\theta) \geq U_\theta(g_{\theta,j}) + \epsilon$$

and taking $j \to \infty$ yields:

$$U_\theta(f_\theta) \geq U_\theta(g_\theta) + \epsilon > U_\theta(g_\theta)$$

as desired. □

Proof of Theorem 4. Necessity of the axioms (the backwards directions of (i), (iii), (v)) is straightforward, so we prove sufficiency. Throughout, we study $\succsim_\theta$ restricted to the domain $\mathcal{F}_0(\mathcal{X})$. Note that this corresponds to the choice environment of Maccheroni et al. (2006) where the state space is $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and the consequence space is $\mathbb{R}$.

Part (i): We show the forwards direction. It is straightforward to show that Axioms 6–12 imply that $\succsim_\theta$ on $\mathcal{F}_0(\mathcal{X})$ satisfies Axioms A.1-A.6 of Maccheroni et al. (2006). By Theorem 3 of Maccheroni et al. (2006), there exist a nonconstant affine function $u_\theta : \mathbb{R} \to \mathbb{R}$ and convex, weak*-lower-semicontinuous function $c_{\theta,u_\theta} : \Delta^F(\mathcal{X}) \to [0, \infty]$ satisfying $\inf_{P \in \Delta^F(\mathcal{X})} c_{\theta,u_\theta}(P) = 0$ such that the function

$$f \mapsto \min_{P \in \Delta^F(\mathcal{X})} \left(\int u_\theta(f) \, dP + c_{\theta,u_\theta}(P) \right)$$

represents $\succsim_\theta$ on $\mathcal{F}_0(\mathcal{X})$. By cardinal uniqueness (Corollary 5 of Maccheroni et al. (2006)) and Monotonicity, we may choose $u_\theta = \mathrm{Id}$ and define $c_\theta = c_{\theta,\mathrm{Id}}$. Hence, the function

$$U_\theta(f) = \min_{P \in \Delta^F(\mathcal{X})} \left(\int f \, dP + c_\theta(P) \right)$$

represents $\succsim_\theta$ on $\mathcal{F}_0(\mathcal{X})$. Conclude by applying Lemmas 1 and 2.

Part (ii): Since $\mathbb{R}$ is unbounded, uniqueness and the given expression immediately follow from Proposition 6 of Maccheroni et al. (2006).

Part (iii): Additionally suppose that $\succsim_\theta$ satisfies Axiom 13. By Lemma 32 of Maccheroni et al. (2006), $c_\theta(Q_\theta) = 0$.

Part (iv): Additionally suppose that $\succsim_\theta$ satisfies Axiom 14. The equalities

$$\{P \in \Delta^F(\mathcal{X}) : c_\theta(P) \leq t\} = \overline{\{P \in \Delta(\mathcal{X}) : c_\theta(P) \leq t\}} \quad \forall t \geq 0$$

immediately follow from Parts (i) and (ii). It remains to show

$$V_\theta(L) = \sup_{P \in \Delta(\mathcal{X})} \left\{ \int L_\theta(x) \, dP(x) - c_\theta(P) \right\} \quad \forall L \in \mathcal{L}$$

or equivalently by Lemma 1,

$$\min_{P\in\Delta^F(\mathcal{X})}\left(\int f_\theta\ dP + c_\theta(P)\right) = U_\theta(f) = \inf_{P\in\Delta(\mathcal{X})}\left\{\int f_\theta(x)\ dP(x) + c_\theta(P)\right\} \quad \forall f\in\mathcal{F}$$

for U_θ as defined in Equation (17). Fix any $f\in\mathcal{F}$. The inequality $\leq$ follows immediately. For the other inequality $\geq$, let $P^*\in\Delta^F(\mathcal{X})$ such that

$$U_\theta(f) = \int f_\theta\ dP^* + c_\theta(P^*)$$

and define $t^* = c_\theta(P^*)$. Since

$$P^*\in\{P\in\Delta^F(\mathcal{X}) : c_\theta(P)\leq t^*\} = \overline{\{P\in\Delta(\mathcal{X}) : c_\theta(P)\leq t^*\}}$$

by Aliprantis and Border (2006) Theorem 2.14, there exists a net

$$(P_\alpha)_{\alpha\in A}\subseteq\{P\in\Delta(\mathcal{X}) : c_\theta(P)\leq t^*\}$$

such that $P_\alpha\to P^*$ in the weak* topology. By definition of weak* lower semicontinuity,

$$\liminf_{\alpha\in A} c_\theta(P_\alpha)\geq t^*$$

By above,

$$\begin{aligned} & c_\theta(P_\alpha)\leq t^* \quad \forall\alpha\in A \\ \implies & \sup_{\alpha\geq\alpha_0} c_\theta(P_\alpha)\leq t^* \quad \forall\alpha_0\in A \\ \implies & \limsup_{\alpha\in A} c_\theta(P_\alpha)\leq t^* \end{aligned}$$

Hence, $c_\theta(P_\alpha)\to t^*$. Since $P\mapsto\int f_\theta\ dP$ is weak* continuous (since f is bounded and measurable), we have by Theorem 2.28 of Aliprantis and Border (2006) that $\int f_\theta\ dP_\alpha\to\int f_\theta\ dP^*$. Finally, note that

$$\begin{aligned} \int f_\theta\ dP_\alpha + c_\theta(P_\alpha) \geq \inf_{P\in\Delta(\mathcal{X})}\left(\int f_\theta\ dP + c_\theta(P)\right) \quad \forall\alpha\in A \\ \implies U_\theta(f)\geq\inf_{P\in\Delta(\mathcal{X})}\left(\int f_\theta\ dP + c_\theta(P)\right) \end{aligned}$$

as desired. The remaining assertion immediately follows by substituting $L = 0$ and noting that $\inf_{P \in \Delta^F(\mathcal{X})} c_\theta(P) = 0$.

Part (v): Additionally suppose that $\succsim_\theta$ satisfies Axiom 15. By Parts (i) and (ii), this implies: for each $t \geq 0$, the set

$$\{P \in \Delta(\mathcal{X}) : c_\theta(P) \leq t\}$$

is weakly closed. By definition, $c_\theta : \Delta(\mathcal{X}) \to [0, \infty]$ is weakly lower-semicontinuous. □

A.3 Axioms for Aggregation Across Θ

While the preferences $\{\succsim_\theta : \theta \in \Theta\}$ describe the researcher's preferences under known θ, θ is unknown in practice, so our ultimate interest is in the overall preference $\succsim_\Theta$. The remaining axioms for $\succsim_\Theta$ relate it to the conditional preferences.

Axiom 16 (Consistency)**.** *If $L \succsim_\theta L'$ for all $\theta \in \Theta$, then $L \succsim_\Theta L'$.*

Consistency requires that if all conditional preferences agree on a ranking, the overall preference must respect that ranking.

Axiom 17 (Caution)**.** *For all $r \in \mathbb{R}$ and $L \in \mathcal{L}$, if there exists $\theta \in \Theta$ such that $r \succ_\theta L$, then $r \succsim_\Theta L$.*

Caution requires that if there exists some $\theta \in \Theta$ under which a constant loss r is strictly preferred to a state-dependent loss L, then the overall preference must also (weakly) prefer r to L. This is an uncertainty-averse aggregation rule: the researcher is cautious about state-dependent losses whenever any parameter value suggests caution.

Definition 3. *The family of preferences $\{\succsim_\Theta\} \cup \{\succsim_\theta : \theta \in \Theta\}$ has a* cautious variational representation *if each $\succsim_\theta$ has a variational representation as defined in Equation (16) and*

$$V_\Theta(L) = \sup_{\theta \in \Theta} V_\theta(L) = \sup_{\theta \in \Theta} \max_{P \in \Delta^F(\mathcal{X})} \left\{ \int L_\theta(x) dP(x) - c_\theta(P) \right\} \tag{18}$$

represents $\succsim_\Theta$, in the sense that $L \succsim_\Theta L'$ if and only if $V_\Theta(L) \leq V_\Theta(L')$.

Theorem 5. *(i) The family of preferences $\{\succsim_\Theta\} \cup \{\succsim_\theta : \theta \in \Theta\}$ satisfies Axioms 6–12 and 16–17 if and only if it has a cautious variational representation.*

(ii) It additionally satisfies Axioms 13-15 if and only if each c_θ satisfies: $c_\theta(Q_\theta) = 0$, $\{c_\theta \leq t\} = \overline{\{c_\theta \leq t\} \cap \Delta(\mathcal{X})}$ for all $t \geq 0$, and $c_\theta : \Delta(\mathcal{X}) \to [0, \infty]$ is weakly lower semicontinuous. In this case, we also have: $\inf_{P \in \Delta(\mathcal{X})} c_\theta(P) = 0$ and, for V_Θ defined in Equation (18),

$$V_\Theta(L) = \sup_{\theta \in \Theta} \sup_{P \in \Delta(\mathcal{X})} \left(\int L_\theta \, dP - c_\theta(P) \right) \quad \forall L \in \mathcal{L}$$

Part (i) of Theorem 5 adapts arguments from Cerreia-Vioglio et al. (2025) to our setting with loss functions as primitives. It shows that, under the axioms above, the preference $\succsim_\Theta$ takes a min-max approach to parameter uncertainty, together with a penalized min-max approach to model misspecification. As we discuss in the text, this class of preferences over loss functions nests some prominent examples which have previously been discussed in the literature on estimation under misspecification.

Proof of Theorem 5. Part (ii) immediately follows from Part (i) and the previous theorem. Furthermore, necessity of the axioms in Part (i) (the backwards direction of Part (i)) is straightforward, so we prove sufficiency.

Part (i): By the exact analog of Lemma 1, V_Θ represents $\succsim_\Theta$ on $\mathcal{L}$ if and only if

$$U_\Theta(f) = \inf_{\theta \in \Theta} U_\theta(f) = \inf_{\theta \in \Theta} \min_{P \in \Delta^F(\mathcal{X})} \left(\int f_\theta(x) \, dP(x) + c_\theta(P) \right) \tag{19}$$

represents $\succsim_\Theta^{\mathcal{F}}$ on $\mathcal{F}$. As before, we abuse notation and refer to $\succsim_\Theta^{\mathcal{F}}$ as $\succsim_\Theta$ henceforth. By Theorem 4, each $\succsim_\theta$ has a variational representation U_θ on $\mathcal{F}$.

Step 1: $\succsim_\Theta$ admits CEs on $\mathcal{F}$. Fix any $f \in \mathcal{F}$. Since f is bounded, there exist $\overline{k} \geq \underline{k} \in \mathbb{R}$ such that $\overline{k} \geq f(\theta, x) \geq \underline{k}$ for all (θ, x). By Monotonicity, $\overline{k} \succsim_\Theta f \succsim_\Theta \underline{k}$. By Θ-Mixture Continuity and Monotonicity, there exists a unique $\alpha \in [0, 1]$ such that $f \sim_\Theta \alpha\overline{k} + (1 - \alpha)\underline{k}$. Hence, the certainty equivalent function $CE_\Theta : \mathcal{F} \to \mathbb{R}$ defined as: $f \sim_\Theta CE_\Theta(f)$ is well-defined. By Monotonicity, $CE_\Theta(\cdot)$ represents $\succsim_\Theta$ on $\mathcal{F}$.

Step 2: $CE_\Theta \geq U_\Theta$. Define U_Θ as in Equation (19). Fix any $f \in \mathcal{F}$ and consider the constant act $U_\Theta(f)$. Note that since $U_\theta(f) \geq U_\Theta(f) = U_\theta(U_\Theta(f))$ for all $\theta \in \Theta$,

$$f \succsim_\theta U_\Theta(f) \quad \forall \theta \in \Theta$$

and hence Consistency implies $f \succsim_\Theta U_\Theta(f)$. Since CE_Θ is normalized, $CE_\Theta(f) \geq U_\Theta(f)$.

Step 3: $U_\Theta \geq CE_\Theta$. Fix any $f \in \mathcal{F}$ and consider the constant act $U_\Theta(f)$. For each $\epsilon > 0$, note that there exists $\theta \in \Theta$ such that

$$U_\Theta(f) + \epsilon > U_\theta(f) \implies U_\Theta(f) + \epsilon \succ_\theta f$$

By Caution and since CE_Θ is normalized,

$$U_\Theta(f) + \epsilon \succsim_\Theta f \implies U_\Theta(f) + \epsilon \geq CE_\Theta(f)$$

Taking $\epsilon \downarrow 0$ yields $U_\Theta(f) \geq CE_\Theta(f)$. We have therefore shown that $U_\Theta = CE_\Theta$ represents $\succsim_\Theta$ on $\mathcal{F}$, as desired. □

B Proofs

B.1 Proofs for Results in Section 2

Constraint Preferences. To characterize constraint preferences, we introduce another axiom. Define

$$\mathcal{P}_\theta := \{P \in \Delta^F(\mathcal{X}) : \succsim_\theta \text{ is more ambiguity averse than } \succsim_P^{SEU}\}$$

Axiom 18. *For all $\theta \in \Theta$, $\mathcal{P}_\theta = \overline{\mathcal{P}_\theta \cap \Delta(\mathcal{X})}$ and $\mathcal{P}_\theta \cap \Delta(\mathcal{X})$ is weakly closed.*

We prove the following result, which implies Propositions 1 and 2.

Theorem 6. *(i) The preference $\succsim_\theta$ satisfies Axioms 6-10, Certainty Independence, and Axiom 12 if and only if $\mathcal{P}_\theta$ is nonempty, convex, and weak*-closed and $\succsim_\theta$ has a* maxmin expected utility *representation with ambiguity set $\mathcal{P}_\theta$: the function*

$$V_\theta(L) = \max_{P \in \mathcal{P}_\theta} \int L_\theta \, dP \tag{20}$$

represents $\succsim_\theta$ on $\mathcal{L}$.

(ii) Suppose $\succsim_\theta$ satisfies Axioms 6-10, Certainty Independence, and Axiom 12. It additionally satisfies Axiom 13 if and only if $Q_\theta \in \mathcal{P}_\theta$.

(iii) Suppose $\succsim_\theta$ satisfies Axioms 6-10, Certainty Independence, and Axiom 12. It additionally satisfies Axiom 18 if and only if it satisfies Axioms 14 and 15. In this case,

$$V_\theta(L) = \sup_{P \in \mathcal{P}_\theta \cap \Delta(\mathcal{X})} \int L_\theta \, dP$$

Proof of Theorem 6. Necessity of the axioms (the backwards directions of each part) is straightforward, so we prove sufficiency.

Part (i): By Lemmas 1 and 2 and Proposition 19(iii) of Maccheroni et al. (2006), c_θ takes only values 0 and ∞. By Lemma 32 of Maccheroni et al. (2006),

$$\mathcal{P}_\theta = \{P \in \Delta^F(\mathcal{X}) : c_\theta(P) = 0\}$$

Since $c_\theta : \Delta^F(\mathcal{X}) \to [0, \infty]$ is weak* lower semicontinuous and convex, $\mathcal{P}_\theta$ is weak* closed and convex. Since $\Delta^F(\mathcal{X})$ is weak* compact, $0 = \inf_{P \in \Delta^F(\mathcal{X})} c_\theta(P) = \min_{P \in \Delta^F(\mathcal{X})} c_\theta(P)$, so $\mathcal{P}_\theta$ is nonempty. Finally, Equation (20) follows from Proposition 19(ii) of Maccheroni et al. (2006).

Part (ii): This immediately follows from Lemma 32 of Maccheroni et al. (2006) and Theorem 4(iii).

Part (iii): Suppose $\succsim_\theta$ satisfies Axioms 6-10, Certainty Independence, and Axiom 12. By Lemma 32 of Maccheroni et al. (2006), for any $t \geq 0$,

$$\mathcal{P}_\theta = \{P \in \Delta^F(\mathcal{X}) : c_\theta(P) = 0\} = \{P \in \Delta^F(\mathcal{X}) : c_\theta(P) \leq t\}$$

Hence, $\succsim_\theta$ satisfies Axiom 18 if and only if it satisfies Axioms 14 and 15. The desired representation then follows from Theorem 4(iv).

Multiplier Preferences. Recall that an event $\mathcal{E} \subseteq \mathcal{X}$ is *nonnull under* $\succsim_\theta$ if there exist $L, L', M \in \mathcal{L}$ such that $L_\mathcal{E} M \succ_\theta L'_\mathcal{E} M$.

Lemma 3. *Under the basic axioms, if $\mathcal{E}$ is nonnull under $\succsim_\theta$, then there exist $L, L', M \in \mathcal{L}_0$ such that $L_\mathcal{E} M \succ_\theta L'_\mathcal{E} M$.*

Proof of Lemma 3. Let $L, L', M \in \mathcal{L}$ such that $L_\mathcal{E} M \succ_\theta L'_\mathcal{E} M$. By θ-Relevance, $(L_\theta)_\mathcal{E} M_\theta \succ_\theta (L'_\theta)_\mathcal{E} M_\theta$. Define $L_{\theta,j}$ and $M_{\theta,j}$ as $g_{\theta,j}$ in the proof of Lemma 2, and define $L'_{\theta,k}$ and $M_{\theta,k}$ as

$f_{\theta,k}$ in the proof of Lemma 2. For each fixed $k \geq 1$, Monotonicity implies:

$$(L_{\theta,k})_{\mathcal{E}}(M_{\theta,k}) \succsim_\theta (L_\theta)_{\mathcal{E}} M_\theta \succ_\theta (L'_\theta)_{\mathcal{E}} M_\theta$$

By θ-Continuity and since $(L'_{\theta,j})_{\mathcal{E}} M_{\theta,j} \to (L'_\theta)_{\mathcal{E}} M_\theta$, there exists $j \geq 1$ large enough such that

$$(L_\theta)_{\mathcal{E}} M_\theta \succ_\theta (L'_{\theta,j})_{\mathcal{E}} M_{\theta,j}$$

Choose $k = j$. □

The following result implies Proposition 3.

Theorem 7. *Throughout this result, assume that for each $\theta \in \Theta$, $\mathcal{X}$ has at least three disjoint events that are nonnull under $\succsim_\theta$.*

(i) Fix any $\theta \in \Theta$. The preference $\succsim_\theta$ satisfies Axioms 6–13, Monotone Continuity, and the Sure Thing Principle if and only if $\succsim_\theta$ has a multiplier representation *with reference probability Q_θ: there exists $\lambda_\theta \in (0, \infty]$ such that the function*

$$V_\theta(L) = \max_{P \in \Delta(\mathcal{X})} \left(\int L_\theta \, dP - \lambda_\theta \cdot KL(P||Q_\theta) \right) \tag{21}$$

represents $\succsim_\theta$ on $\mathcal{L}$.

(ii) Suppose each $\succsim_\theta$ satisfies the axioms from Part (i). Assume that there exist a collection of events $\{E_\theta\}_{\theta \in \Theta}$ and $q \in (0,1)$ such that $Q_\theta(\mathcal{E}_\theta) = q$ for all $\theta \in \Theta$. The family of preferences additionally satisfies Uniform Misspecification Concern if and only if λ_θ is constant across $\theta \in \Theta$.

Proof of Theorem 7. Necessity of the axioms (the backwards directions of each part) are straightforward, so we prove sufficiency.

Part (i): Fix any $\theta \in \Theta$. By Lemmas 1-3 and Theorem 1 of Strzalecki (2011),[13] there exist $\lambda_\theta \in (0, \infty]$ and $\tilde{Q}_\theta \in \Delta(\mathcal{X})$ such that

$$\max_{P \in \Delta(\mathcal{X})} \left(\int L_\theta \, dP - \lambda_\theta \cdot KL(P||\tilde{Q}_\theta) \right)$$

[13] More precisely, by the portion of the argument for the proof of Theorem 1 of Strzalecki (2011) which delivers the representation for utility acts.

represents $\succsim_\theta$ on $\mathcal{L}$. By Axiom 13 and uniqueness of c_θ, $\lambda_\theta \cdot KL(Q_\theta||\tilde{Q}_\theta) = 0$ and hence $\tilde{Q}_\theta = Q_\theta$, which delivers Equation (21).

Part (ii): Suppose that there exist θ, θ' such that $\lambda_\theta > \lambda_{\theta'}$. By Lemma 1, it suffices to show that the induced preferences over utility acts do not satisfy Uniform Misspecification Concern. Define $f_\theta = 1_{\mathcal{E}_\theta} 0$ and $g_{\theta'} = 1_{\mathcal{E}_{\theta'}} 0$. Note that the law of f_θ under Q_θ coincides with the law of $g_{\theta'}$ under $Q_{\theta'}$: $q\delta_1 + (1-q)\delta_0$. However, by the dual representation of multiplier preferences, we have for $\phi_\lambda(x) = -\exp(-\lambda^{-1}x)$:

$$U_\theta(f) = \phi_{\lambda_\theta}^{-1}\left(\int_S \phi_{\lambda_\theta}(f_\theta)\ dQ_\theta\right) = \phi_{\lambda_\theta}^{-1}\left(q\phi_{\lambda_\theta}(1) + (1-q)(-1)\right)$$

and similarly,

$$U_{\theta'}(g) = \phi_{\lambda_{\theta'}}^{-1}\left(q\phi_{\lambda_{\theta'}}(1) + (1-q)(-1)\right)$$

It is straightforward to show that, for any $q \in (0,1)$, the map

$$\lambda \mapsto \phi_\lambda^{-1}\left(q\phi_\lambda(1) + (1-q)(-1)\right) = -\lambda \log\left(q\exp(-\lambda^{-1}) + (1-q)\right)$$

is strictly increasing in λ, and hence

$$U_{\theta'}(g_{\theta'}) < U_\theta(f_\theta)$$

Note that U_θ and $U_{\theta'}$ are normalized. Hence, choosing $k \in (U_{\theta'}(g_{\theta'}), U_\theta(f_\theta))$ yields a violation of Uniform Misspecification Concern, since

$$k \succ_{\theta'} g_{\theta'} \quad \text{and} \quad f_\theta \succ_\theta k$$

□

Constrained Multiplier We show the following result, which implies Theorem 1.

Theorem 8. *Fix any $\theta \in \Theta$. Suppose that for each $i = 1, 2$, $\succsim_{\theta,i}$ has a variational representation on $\mathcal{L}$: there exists a convex, weak* lower-semicontinuous function $c_{\theta,i} : \Delta^F(\mathcal{X}) \to [0,\infty]$ with $\inf_{P\in\Delta^F(\mathcal{X})} c_{\theta,i}(P) = 0$ such that*

$$V_{\theta,i}(L) = \max_{P\in\Delta^F(\mathcal{X})}\left(\int L_\theta\ dP - c_{\theta,i}(P)\right) \tag{22}$$

represents $\succsim_{\theta,i}$ *on* $\mathcal{L}$*, and additionally suppose that* $c_{\theta,i}(Q_\theta) = 0$ *for each* $i = 1, 2$*.* $\succsim_\theta$ *satisfies the Basic Axioms (*θ*-Relevance, Nontrivial Weak Order, Monotonicity,* θ*-Continuity) and* $(\succsim_\theta, \succsim_{\theta,1}, \succsim_{\theta,2})$ *satisfy Indirect Pareto if and only if* $\succsim_\theta$ *has a variational representation*

$$V_\theta(L) = \max_{P \in \Delta^F(\mathcal{X})} \left(\int L_\theta \, dP - (c_{\theta,1}(P) + c_{\theta,2}(P)) \right) \tag{23}$$

on $\mathcal{L}$*. In particular for the case where* $c_{\theta,1}(P) = \lambda \cdot KL(P||Q_\theta)$ *and* $c_{\theta,2} = \chi_{\mathcal{P}_\theta}(P)$ *for some nonempty, convex, weak* closed* $\mathcal{P}_\theta \subseteq \Delta^F(\mathcal{X})$ *containing* Q_θ*, we may write Equation 23 as:*

$$V_\theta(L) = \sup_{P \in \Delta(\mathcal{X}) \cap \mathcal{P}_\theta} \left(\int L_\theta \, dP - \lambda KL(P||Q_\theta) \right) \tag{24}$$

The following definitions and lemmas will be useful for the proof of Theorem 8. Let V be a real vector space. The *strict epigraph* of a function $F : V \to [-\infty, +\infty]$ is the set $\text{epi}_S(F) := \{(v, r) \in V \times \mathbb{R} : F(v) < r\}$.

Definition 4 (Zălinescu (2002) Theorem 2.1.3(ix))**.** *Given functions* $F, G : V \to (-\infty, +\infty]$*, their* inf-convolution *is the function* $F \square G : V \to [-\infty, +\infty]$*, where:*

$$(F \square G)(v) = \inf_{v_1, v_2 \in V : v_1 + v_2 = v} \Big(F(v_1) + G(v_2) \Big) \tag{25}$$

Lemma 4. *For any* $F, G : V \to (-\infty, +\infty]$*,* $epi_S(F \square G) = epi_S(F) + epi_S(G)$*.*[14]

Proof. First, we show the forwards inclusion: $\text{epi}_S(F \square G) \subseteq \text{epi}_S(F) + \text{epi}_S(G)$. Note that

$$\begin{aligned} & (v, r) \in \text{epi}_S(F \square G) \\ \iff & [F \square G](v) = \inf_{v_1 + v_2 = v} \Big(F(v_1) + G(v_2) \Big) < r \\ \iff & \exists \; y_1 + y_2 = v \text{ s.t. } F(y_1) + G(y_2) < r \end{aligned}$$

Let $\alpha := \frac{1}{2}(r - F(y_1) - G(y_2)) > 0$, and define $r_1 := F(y_1) + \alpha$ and $r_2 := G(y_2) + \alpha$. Then we have $(y_1, r_1) \in \text{epi}_S(F)$ and $(y_2, r_2) \in \text{epi}_S(G)$ with $(y_1, r_1) + (y_2, r_2) = (v, r)$.

Second, we show the backwards inclusion: $\text{epi}_S(F) + \text{epi}_S(G) \subseteq \text{epi}_S(F \square G)$. Let $(y_1, r_1) \in$

[14] This relation is stated without proof as Equation (2.6) of Zălinescu (2002). For completeness, we provide a proof here, which follows immediately from the definitions.

$\text{epi}_S(F)$ and $(y_2, r_2) \in \text{epi}_S(G)$. Note that by definition,

$$[F \square G](y_1 + y_2) = \inf_{v_1 + v_2 = y_1 + y_2} \Big(F(v_1) + G(v_2) \Big) \leq F(y_1) + G(y_2) < r_1 + r_2$$

and hence $(y_1 + y_2, r_1 + r_2) \in \text{epi}_S(F \square G)$. □

Let $\mathcal{L}(\mathcal{X})$ be the Banach space of bounded, Borel measurable functions $L : \mathcal{X} \to \mathbb{R}$, endowed with the sup norm.[15] Let $ba(\mathcal{X})$ be the real vector space of bounded, finitely additive, signed Borel measures on $\mathcal{X}$. By Aliprantis and Border (2006) Theorem 14.4, $ba(\mathcal{X})$ is the topological (norm) dual of $\mathcal{L}(\mathcal{X})$. Recall that we endowed $ba(\mathcal{X})$ (and $\Delta^F(\mathcal{X})$) with the corresponding weak*-topology. Note that under this topology, $ba(\mathcal{X})$ is a separated, locally convex vector space whose topological dual is $\mathcal{L}(\mathcal{X})$.

Lemma 5. *For each $i = 1, 2$, let $c_i : \Delta^F(\mathcal{X}) \to [0, \infty]$ be a convex, weak* lower-semicontinuous function with $\inf_{P \in \Delta^F(\mathcal{X})} c_i(P) = 0$, let $V_i : \mathcal{L}(\mathcal{X}) \to \mathbb{R}$ be defined as in Equation (22):*

$$V_i(L) = \max_{P \in \Delta^F(\mathcal{X})} \left(\int L \, dP - c_i(P) \right)$$

and let $V : \mathcal{L}(\mathcal{X}) \to \mathbb{R}$ be defined as in Equation (23):

$$V(L) = \max_{P \in \Delta^F(\mathcal{X})} \left(\int L \, dP - c_1(P) - c_2(P) \right)$$

Assume that $c_1(Q) = c_2(Q) = 0$ for some $Q \in \Delta^F(\mathcal{X})$. Then, $V = V_1 \square V_2$.

Proof. For each $i = 1, 2$, define $\overline{c}_i : ba(\mathcal{X}) \to [0, \infty]$ as:

$$\overline{c}_i(P) = \begin{cases} c_i(P) & P \in \Delta^F(\mathcal{X}) \\ +\infty & \text{else} \end{cases} \tag{26}$$

By definition of conjugate (Equation (2.30) of Zălinescu (2002)), $V_i = \overline{c}_i^*$. It is straightforward to verify that $\overline{c}_i$ is proper, convex, and weak* lower-semicontinuous. By the Fenchel–Moreau theorem (Theorem 2.3.3 of Zălinescu (2002)), $V_i^* = \overline{c}_i^{**} = \overline{c}_i$. By Theorem 2.3.1 of Zălinescu (2002), $(V_1 \square V_2)^* = V_1^* + V_2^* = \overline{c}_1 + \overline{c}_2$. Since each V_i is convex and proper, $V_1 \square V_2$ is convex by Theorem 2.1.3 of Zălinescu (2002).

[15] In particular, $\mathcal{L}(\mathcal{X})$ is a separated, locally convex vector space, as in the setup of Zălinescu (2002) Section 2.3.

For ease of notation, let $W = V_1 \square V_2$. Next, we show that W is 1-Lipschitz, and hence lower-semicontinuous. First, we recall that since each V_i is 1-Lipschitz, for any $L'', H \in \mathcal{L}$,

$$V_2(L'' + H) - V_2(L'') \leq |V_2(L'' + H) - V_2(L'')| \leq \|H\|_\infty$$

Hence, for any $L, L', L'', H \in \mathcal{L}$ with $L = L' + L''$,

$$W(L + H) \leq V_1(L') + V_2(L'' + H) \leq V_1(L') + V_2(L'') + \|H\|_\infty$$

Taking the inf over $L', L'' \in \mathcal{L}$ with $L' + L'' = L$ yields:

$$W(L + H) \leq W(L) + \|H\|_\infty$$

An analogous argument shows that $W(L) \leq W(L + H) + \|H\|_\infty$. Hence, W is 1-Lipschitz. Finally, we show that W is proper. By definition, $W < +\infty$. By assumption, for any $L, L' \in \mathcal{L}$, we have $V_1(L') \geq \int L' \, dQ$ and $V_2(L - L') \geq \int (L - L') \, dQ$. Hence,

$$W(L) \geq \int L \, dQ > -\infty$$

Since W is proper, convex, and lower semicontinuous, another application of the Fenchel–Moreau theorem yields: $W^{**} = W$. Taking conjugates of both sides of $W^* = \overline{c}_1 + \overline{c}_2$ yields

$$V_1 \square V_2 = (\overline{c}_1 + \overline{c}_2)^* = V$$

as desired. □

Lemma 6. *Assume the assumptions of Theorem 8.* $(\succsim_\theta, \succsim_{\theta,1}, \succsim_{\theta,2})$ *satisfy Indirect Pareto if and only if: for each* $L \in \mathcal{L}$ *and* $r \in \mathbb{R}$,

$$L \succ_\theta r \iff V_\theta(L_\theta) < r$$

where V_θ *is defined as in Equation (23).*

Proof. Fix any $i = 1, 2$. Note that: for each $L \in \mathcal{L}$ and $r \in \mathbb{R}$,

$$L \succ_{\theta,i} r \iff V_{\theta,i}(L_\theta) < r \iff (L_\theta, r) \in \mathrm{epi}_S(V_{\theta,i})$$

Hence, $(\succsim_\theta, \succsim_{\theta,1}, \succsim_{\theta,2})$ satisfy Indirect Pareto if and only if: for each $L \in \mathcal{L}$ and $r \in \mathbb{R}$,

$$\begin{aligned} L \succ_\theta r &\iff \\ \exists L_1, L_2 \in \mathcal{L},\ r_1, r_2 \in \mathbb{R} \text{ s.t. } L = L_1 + L_2,\ r = r_1 + r_2,\ (L_{i\theta}, r_i) \in \mathrm{epi}_S(V_{\theta,i}) &\\ \iff (L_\theta, r) \in \mathrm{epi}_S(V_{\theta,1}) + \mathrm{epi}_S(V_{\theta,2}) = \mathrm{epi}_S(V_{\theta,1} \square V_{\theta,2}) = \mathrm{epi}_S(V_\theta) &\iff V_\theta(L_\theta) < r \end{aligned}$$

where the last two equalities follow from the previous lemmas. □

Proof of Theorem 8. Necessity of Indirect Pareto immediately follows from Lemma 6, and necessity of the remaining axioms are straightforward. To prove sufficiency, observe that under the Basic Axioms,

$$L \succsim_\theta L' \iff L_\theta \succsim_\theta L'_\theta \iff \inf\{r : L_\theta \succ_\theta r\} \leq \inf\{r : L'_\theta \succ_\theta r\}$$

where the first equivalence follows from θ-Relevance. For the forwards direction, observe that if $L_\theta \succsim_\theta L'_\theta$ $\{r : L'_\theta \succ_\theta r\} \subseteq \{r : L_\theta \succ_\theta r\}$. For the backwards direction, assume that $L_\theta \succ_\theta L'_\theta$. By θ-Continuity, there exists $r, r' \in \mathbb{R}$ such that $L_\theta \succ_\theta r \succ_\theta r' \succ_\theta L'_\theta$. Hence,

$$\inf\{r : L_\theta \succ_\theta r\} \leq r < r' \leq \inf\{r : L'_\theta \succ_\theta r\}$$

By Lemma 6, $\inf\{r : L_\theta \succ_\theta r\} = V_\theta(L_\theta)$. Finally, to prove the representation Equation (24) in the case where $c_{\theta,1}(P) = \lambda \cdot KL(P||Q_\theta)$ and $c_{\theta,2} = \chi_{\mathcal{P}_\theta}(P)$ for some nonempty, convex, weak* closed $\mathcal{P}_\theta \subseteq \Delta^F(\mathcal{X})$ containing Q_θ, define $c_\theta = c_{\theta,1} + c_{\theta,2}$. For each $t \geq 0$, define

$$\{c_\theta \leq t\} = \{P \in \Delta^F(\mathcal{X}) : c_\theta(P) \leq t\}$$

By definition of KL divergence,

$$\{c_\theta \leq t\} = \{P \in \Delta(\mathcal{X}) : \lambda KL(P||Q_\theta) \leq t\} \cap \mathcal{P}_\theta \subseteq \Delta(\mathcal{X})$$

Hence,

$$\begin{aligned} \{c_\theta \leq t\} &= \{c_\theta \leq t\} \cap \Delta(\mathcal{X}) \\ \implies \{c_\theta \leq t\} = \overline{\{c_\theta \leq t\}} &= \overline{\{c_\theta \leq t\} \cap \Delta(\mathcal{X})} \end{aligned}$$

Equation (24) then follows from Theorem 4(iv). □

Proof of Proposition 5 The primal problem is

$$V_\theta(L) = \sup_{P \in \mathcal{P}_\theta} \left\{ \int L(\theta, x)\, dP(x) - \lambda \cdot \mathrm{KL}(P \| Q_\theta) \right\}$$

where $\mathcal{P}_\theta = \{P : E_P[\varphi(\theta, X)] = 0\}$. Factoring out λ,

$$V_\theta(L) = -\lambda \inf_{P: E_P[\varphi]=0} \left\{ \mathrm{KL}(P \| Q_\theta) + \int \left(-\frac{1}{\lambda} L(\theta, x) \right) dP(x) \right\}.$$

Define the tilted distribution P_0 by

$$\frac{dP_0}{dQ_\theta}(x) = \frac{\exp\left(\frac{1}{\lambda} L(\theta, x)\right)}{E_{Q_\theta}\left[\exp\left(\frac{1}{\lambda} L(\theta, X)\right)\right]}.$$

A direct calculation shows that

$$\mathrm{KL}(P \| Q_\theta) - \frac{1}{\lambda} \int L(\theta, x)\, dP(x) = \mathrm{KL}(P \| P_0) - \log E_{Q_\theta}\left[\exp\left(\frac{1}{\lambda} L(\theta, X) \right) \right],$$

so

$$V_\theta(L) = \lambda \log E_{Q_\theta}\left[\exp\left(\frac{1}{\lambda} L(\theta, X) \right) \right] - \lambda \inf_{P: E_P[\varphi]=0} \mathrm{KL}(P \| P_0).$$

For the applications below we invoke this result for L such that the right hand side is finite, since in the case where it is infinite $V_\theta(L)$ is infinite as well. The remaining step is a standard KL minimization argument for linear moment constraints; see, for example, Kitamura (2009), which implies that

$$\inf_{P: E_P[\varphi]=0} \mathrm{KL}(P \| P_0) = \sup_{\beta \in \mathbb{R}^b} \left\{ -\log E_{P_0}\left[\exp\left(\beta' \varphi(\theta, X) \right) \right] \right\}.$$

Since

$$E_{P_0}\left[\exp\left(\beta' \varphi(\theta, X) \right) \right] = \frac{E_{Q_\theta}\left[\exp\left(\frac{1}{\lambda} L(\theta, X) + \beta' \varphi(\theta, X) \right) \right]}{E_{Q_\theta}\left[\exp\left(\frac{1}{\lambda} L(\theta, X) \right) \right]},$$

substituting and simplifying yields

$$V_\theta(L) = \inf_{\beta \in \mathbb{R}^b} \lambda \cdot \log\left(E_{Q_\theta}\left[\exp\left(\frac{1}{\lambda} L(\theta, X) - \beta' \varphi(\theta, X) \right) \right] \right),$$

where the sign change on β absorbs the negation. □

B.2 Proofs for Results in Section 3

We first state and prove some auxiliary results which will be useful in the proof of Theorem 2.

Proposition 11. *Assume that the model* $\{Q_{n,\theta} : \theta \in \Theta\}$ *is locally asymptotically normal at* $\theta_0 \in \text{int}(\Theta)$ *with scaling coefficient* $\sqrt{n}$ *and nonsingular Fisher information* I_0*. Let*

$$Y_{n,h} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\theta_{n,h}, X_i).$$

Under Assumption 1, if T_n *is a sequence of* $\mathbb{R}^d$*-valued statistics that is tight under* $Q_{n,0}$*, then for every subsequence there exists a further subsequence* $\{s\}$ *and a possibly randomized measurable function* $t : \mathbb{R}^p \times \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ *such that, with* $U \sim \text{Unif}[0,1]$ *independent of* (X, Y) *in the limit experiment* (12) *and* $T = t(X, Y, U)$*,*

$$(T_s, Y_{s,h}) \xrightarrow{d} (T,\, Y + \Psi h) \quad \textit{under } Q_{s,h}$$

for every $h \in \mathbb{R}^p$*.*

Proof of Proposition 11 Let $\{n_j\}$ be any subsequence. Since T_n is tight under $Q_{n,0}$ and $(S_n, Y_{n,0})$ converges jointly by Assumption 1, Prohorov's theorem yields a further subsequence $\{s\} \subseteq \{n_j\}$ along which

$$(T_s, S_s, Y_{s,0}) \xrightarrow{d} (T, S, Y) \quad \text{under } Q_{s,0}.$$

By LAN and Le Cam's third lemma, for every $h \in \mathbb{R}^p$ the same subsequence converges under $Q_{s,h}$ to a limit law that we again denote by (T, S, Y), where the marginal law of (S, Y) is

$$\begin{pmatrix} S \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} I_0 h \\ -\Psi h \end{pmatrix}, \begin{pmatrix} I_0 & -\Psi' \\ -\Psi & \Omega \end{pmatrix} \right).$$

Let $X = I_0^{-1} S$. Then (X, Y) has distribution Q_h in (12). By the standard representation for randomized procedures in the limit experiment (see e.g. Lemma 7.11 in van der Vaart 1998), there exists a function $t : \mathbb{R}^p \times \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ and an independent $U \sim \text{Unif}[0,1]$ such that

$$T = t(X, Y, U)$$

under the limit law. Finally, by contiguity

$$Y_{s,h} = Y_s + \Psi h \xrightarrow{d} Y + \Psi h$$

under $Q_{s,h}$, and the claim follows. □

Lemma 7. *Let $W_{M,h} = w(Y + \Psi h)$ be as in (13), and let A be a non-negative random variable with $A \geq 1$ almost surely. Then the function*

$$G_h(\beta) = E_{Q_h}\left[A \exp\left(\beta' W_{M,h}\right)\right], \qquad \beta \in \mathbb{R}^b,$$

is strictly convex on its effective domain and has compact sublevel sets.

Proof of Lemma 7 Let $v \in \mathbb{R}^b$ be nonzero. Since Ω has full rank, $Y + \Psi h$ has full support on $\mathbb{R}^k$ under Q_h. The random variable $v' W_{M,h}$ is therefore a nontrivial polynomial in $Y + \Psi h$, and is almost surely nonzero. Hence

$$v' \nabla^2 G_h(\beta) v = E_{Q_h}\left[A \exp\left(\beta' W_{M,h}\right) \left(v' W_{M,h}\right)^2\right] > 0$$

whenever $G_h(\beta) < \infty$, so G_h is strictly convex on its effective domain.

To prove compactness of sublevel sets, let $\{\beta_n\}$ satisfy $\|\beta_n\| \to \infty$. After passing to a subsequence, write $\beta_n = \|\beta_n\| u_n$ with $u_n \to u$ and $\|u\| = 1$. Then $u' W_{M,h}$ is again a nonzero polynomial, and because each component of $W_{M,h}$ is centered under Q_h, $u' W_{M,h}$ has mean zero and hence takes strictly positive values on a nonempty open set of Y values. Therefore there exist an open set $B \subset \mathbb{R}^k$ and $c > 0$ such that $u' w(y) \geq 2c$ for all $y \in B$. By continuity of w, for all sufficiently large n we have $u_n' w(y) \geq c$ on B. Since $A \geq 1$ and $Q_h\{Y + \Psi h \in B\} > 0$,

$$G_h(\beta_n) \geq E_{Q_h}\left[\exp\left(\beta_n' W_{M,h}\right) \mathbf{1}\{Y + \Psi h \in B\}\right] \geq Q_h\{Y + \Psi h \in B\} \exp(c\|\beta_n\|) \to \infty.$$

Thus every sublevel set of G_h is bounded. Since G_h is lower semicontinuous, its sublevel sets are thus compact. □

Lemma 8. *Suppose that for some $h \in \mathbb{R}^p$ and some subsequence $\{s\}$,*

$$(T_{s,h}, W_{M,s,h}) \to_d (T_h, W_{M,h}) \quad \textit{under } Q_{s,h},$$

where $T_{s,h} \in \mathbb{R}^d$ *and* $W_{M,s,h} \in \mathbb{R}^b$. *Then*

$$\liminf_{s\to\infty} \inf_{\beta\in\mathbb{R}^b} E_{Q_{s,h}}\left[\ell^*(T_{s,h})\exp(\beta' W_{M,s,h})\right] \geq \inf_{\beta\in\mathbb{R}^b} E_{Q_h}\left[\ell^*(T_h)\exp(\beta' W_{M,h})\right].$$

Proof of Lemma 8 For each s and $\beta \in \mathbb{R}^b$, define

$$F_s(\beta) = E_{Q_{s,h}}\left[\ell^*(T_{s,h})\exp(\beta' W_{M,s,h})\right],$$

and define the limit objective

$$H(\beta) = E_{Q_h}\left[\ell^*(T_h)\exp(\beta' W_{M,h})\right].$$

Let

$$m = \inf_{\beta\in\mathbb{R}^b} H(\beta) \in [0,\infty].$$

We want to show that

$$\liminf_{s\to\infty} \inf_{\beta\in\mathbb{R}^b} F_s(\beta) \geq m.$$

Let $g_r : \mathbb{R}^d \to \mathbb{R}_+$ be bounded Lipschitz functions such that $g_r \uparrow \ell^*$ pointwise, where such a sequence exists by Baire's Theorem (see e.g. Theorem 6.4.1 in Cobasz et al. 2019), and let $\chi_r : \mathbb{R}^b \to [0,1]$ be Lipschitz functions such that $\chi_r \uparrow 1$ pointwise, $\chi_r(w) = 1$ for $\|w\| \leq r$, and $\chi_r(w) = 0$ for $\|w\| \geq r+1$. For $\beta \in \mathbb{R}^b$ define

$$f_r(t,w;\beta) = g_r(t)\chi_r(w)\exp(\beta' w).$$

Then for each fixed β,

$$0 \leq f_r(t,w;\beta) \uparrow \ell^*(t)\exp(\beta' w) \qquad \text{pointwise in } (t,w).$$

Hence, by the monotone convergence theorem,

$$H_r(\beta) = E_{Q_h}[f_r(T_h, W_{M,h};\beta)] \uparrow H(\beta) \qquad \text{for each } \beta.$$

We now argue by contradiction. Suppose first that $m < \infty$ and result fails. Then there exist $\varepsilon > 0$, a further subsequence, again denoted $\{s\}$, and vectors $\beta_s \in \mathbb{R}^b$ such that

$$F_s(\beta_s) \leq m - \varepsilon \qquad \forall s.$$

Suppose next that $m = \infty$ and that the conclusion fails. Then there exist a further subsequence, again denoted $\{s\}$, a finite constant M, and vectors $\beta_s \in \mathbb{R}^b$ such that

$$F_s(\beta_s) \leq M \qquad \forall s.$$

Thus, in either case, there exists a finite constant C such that along a subsequence

$$F_s(\beta_s) \leq C \qquad \forall s.$$

We consider two cases.

Case 1: $\{\beta_s\}$ is bounded. Passing to a further subsequence if necessary, let $\beta_s \to \beta$. Fix r. Since $\{\beta_s\} \cup \{\beta\}$ is contained in a compact set, and $\chi_r(w) = 0$ for $\|w\| \geq r + 1$, the function $f_r(t, w; \tilde{\beta})$ is bounded and continuous on $\mathbb{R}^d \times \mathbb{R}^b \times K$ for any compact set $K \supset \{\beta_s : s \geq 1\} \cup \{\beta\}$. Because

$$(T_{s,h}, W_{M,s,h}, \beta_s) \to_d (T_h, W_{M,h}, \beta),$$

it follows that

$$E_{Q_{s,h}}\left[f_r(T_{s,h}, W_{M,s,h}; \beta_s)\right] \to H_r(\beta).$$

Since

$$f_r(t, w; \beta_s) \leq \ell^*(t) \exp(\beta_s' w) \qquad \text{pointwise},$$

we have

$$E_{Q_{s,h}}\left[f_r(T_{s,h}, W_{M,s,h}; \beta_s)\right] \leq F_s(\beta_s) \leq C \qquad \forall s.$$

Therefore

$$H_r(\beta) \leq C \qquad \forall r.$$

Letting $r \to \infty$ and using monotone convergence gives

$$H(\beta) \leq C.$$

If $m < \infty$, then $C < m$, which contradicts the definition of m. If $m = \infty$, then $C < \infty$, which again contradicts the definition of m.

Case 2: $\|\beta_s\| \to \infty$. Write

$$\beta_s = c_s u_s, \qquad c_s = \|\beta_s\|, \qquad \|u_s\| = 1.$$

Passing to a further subsequence if necessary, let $u_s \to u$ with $\|u\| = 1$. Since Ω has full rank, the Gaussian vector $Y + \Psi h$ has full support on $\mathbb{R}^k$ under Q_h. Hence $u'W_{M,h}$ is a nonzero centered (i.e. mean zero) polynomial in $Y + \Psi h$, and therefore it is not almost surely zero. Since it is centered, it follows that

$$Q_h\{u'W_{M,h} > 0\} > 0.$$

Choose $\eta, \delta > 0$ such that

$$Q_h\{u'W_{M,h} > \eta\} > 2\delta.$$

Because

$$(W_{M,s,h}, u_s) \to_d (W_{M,h}, u),$$

the continuous mapping theorem implies that

$$u_s'W_{M,s,h} \to_d u'W_{M,h}.$$

Therefore, by the Portmanteau theorem, for all sufficiently large s,

$$Q_{s,h}\{u_s'W_{M,s,h} > \eta\} \geq \delta.$$

Since $\ell \geq 0$, we have $\ell^* \geq 1$, so for all sufficiently large s,

$$F_s(\beta_s) \geq E_{Q_{s,h}}\left[\exp(c_s u_s'W_{M,s,h}) 1\{u_s'W_{M,s,h} > \eta\}\right] \geq \delta e^{c_s\eta}.$$

Because $c_s \to \infty$, the right-hand side diverges to ∞, which contradicts $F_s(\beta_s) \leq C$.

Both cases lead to a contradiction. Hence

$$\liminf_{s\to\infty} \inf_{\beta\in\mathbb{R}^b} F_s(\beta) \geq m = \inf_{\beta\in\mathbb{R}^b} H(\beta),$$

as claimed. □

Proof of Theorem 2 For $h \in \mathbb{R}^p$, define $T_{n,h} = \sqrt{n}\left(\delta_n(X^n) - \kappa(\theta_{n,h})\right)$. If $\{T_{n,0}\}$ is not tight under $Q_{n,0}$, then because $Q_{n,0} \in \mathcal{P}_{n,0}^M$ and $\ell(u) \to \infty$ as $\|u\| \to \infty$,

$$\sup_I \liminf_{n\to\infty} \sup_{h\in I} \sup_{P\in\mathcal{P}_{n,h}^M} \{E_P[L_n(\delta_n(X^n), \theta_{n,h})] - \lambda\,\mathrm{KL}(P\|Q_{n,h})\} \geq \liminf_{n\to\infty} E_{Q_{n,0}}[\ell(T_{n,0})] = \infty.$$

Hence the result is trivial in this case. We may therefore restrict attention to subsequences along which $\{T_{n,0}\}$ is tight.

Fix such a subsequence, and let $T_n = T_{n,0} = \sqrt{n}\,(\delta_n(X^n) - \kappa(\theta_0))$. By Proposition 11, there is a further subsequence $\{s\}$ and a possibly randomized statistic $T = t(X, Y, U)$ in the limit experiment such that for every $h \in \mathbb{R}^p$,

$$(T_s, Y_{s,h}) \xrightarrow{d} (T,\ Y + \Psi h) \quad \text{under } Q_{s,h}.$$

Since $T_{s,h} = T_s - \sqrt{s}\,(\kappa(\theta_{s,h}) - \kappa(\theta_0))$, differentiability of κ at θ_0 implies $T_{s,h} \xrightarrow{d} T - Kh$ under $Q_{s,h}$. Hence, by the assumed convergence of moments up to order M and the continuous mapping theorem,

$$(T_{s,h}, W_{M,s,h}) \xrightarrow{d} (T - Kh,\ W_{M,h}) \quad \text{under } Q_{s,h}$$

for every $h \in \mathbb{R}^p$.

Let $\mathbb{Q}^p = \{q_1, q_2, \ldots\}$ and define $I_b = \{q_1, \ldots, q_b\}$. By Lemma 8, for every rational $h \in \mathbb{Q}^p$,

$$\liminf_{s\to\infty} \inf_{\beta} E_{Q_{s,h}} \left[\ell^*(T_{s,h}) \exp(\beta' W_{M,s,h})\right] \geq \inf_{\beta} E_{Q_h} \left[\ell^*(T - Kh) \exp(\beta' W_{M,h})\right].$$

It follows that

$$\sup_{I} \liminf_{n\to\infty} \sup_{h\in I} \inf_{\beta} \lambda \log E_{Q_{n,h}} \left[\ell^*(T_{n,h}) \exp(\beta' W_{M,n,h})\right] \geq$$

$$\lim_{b\to\infty} \liminf_{n\to\infty} \sup_{h\in I_b} \inf_{\beta} \lambda \log E_{Q_{n,h}} \left[\ell^*(T_{n,h}) \exp(\beta' W_{M,n,h})\right] \geq$$

$$\sup_{h\in\mathbb{Q}^p} \inf_{\beta} \lambda \log E_{Q_h} \left[\ell^*(T - Kh) \exp(\beta' W_{M,h})\right].$$

We next show that the right-hand side is unchanged if $\mathbb{Q}^p$ is replaced by $\mathbb{R}^p$. Write

$$G(h, \beta) = E_{Q_h} \left[\ell^*(T - Kh) \exp(\beta' W_{M,h})\right].$$

Using the likelihood ratio of Q_h to Q_0, we may write

$$G(h, \beta) = E_{Q_0} \left[\ell^*(T - Kh) \exp(\beta' W_{M,h}) \exp\left(h' I_0 X - \frac{1}{2} h' I_0 h\right)\right].$$

For each realization (T, X, Y) the integrand is non-negative and continuous in (h, β), so Fatou's lemma implies that G is jointly lower semicontinuous. Moreover, since $\ell^* \geq 1$,

$$G(h, \beta) \geq E_{Q_h} \left[\exp(\beta' W_{M,h}) \right] = E_{Q_0} \left[\exp(\beta' W_{M,0}) \right].$$

By Lemma 7, the right-hand side has compact sublevel sets. Hence Berge's theorem implies that

$$h \mapsto \inf_{\beta} G(h, \beta)$$

is lower semicontinuous. Since $\mathbb{Q}^p$ is dense in $\mathbb{R}^p$, we obtain

$$\sup_{h \in \mathbb{Q}^p} \inf_{\beta} G(h, \beta) = \sup_{h \in \mathbb{R}^p} \inf_{\beta} G(h, \beta).$$

Finally, the statistic $T = t(X, Y, U)$ may depend on the auxiliary randomization variable U. For each fixed h and β, the weight $\exp(\beta' W_{M,h})$ depends only on (X, Y), while ℓ^* is convex because ℓ is convex. Therefore Jensen's inequality implies

$$E_{Q_h} \left[\ell^*(T - Kh) \exp(\beta' W_{M,h}) \right] \geq E_{Q_h} \left[\ell^*(E[T \mid X, Y] - Kh) \exp(\beta' W_{M,h}) \right].$$

Thus randomization cannot improve the criterion, and the lower bound is further bounded below by the infimum over non-randomized decision rules $\delta : \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^d$. This shows that for any sequence of sample sizes, we can extract a further subsequence along which the asserted lower bound holds, which proves the theorem. □

Proof of Corollary 1 By Theorem 2,

$$\inf_{M} \sup_{I} \liminf_{n \to \infty} \sup_{h \in I} \sup_{P \in \mathcal{P}^M_{n,h}} \left\{ \mathbb{E}_P \left[L_n(\delta(X^n), \theta_{n,h}) \right] - \lambda \, \mathrm{KL}(P \| Q_{n,h}) \right\}$$

$$\geq \inf_{M} \inf_{\delta} \sup_{h \in \mathbb{R}^p} \inf_{\beta} \lambda \cdot \log \left(E_{Q_h} \left[\ell^* \left(\delta \left(X, Y \right) - Kh \right) \exp \left(\beta' W_{M,h} \right) \right] \right).$$

By duality,

$$\inf_{\beta} \lambda \log \left(E_{Q_h} \left[\ell^* \left(\delta(X, Y) - Kh \right) \exp \left(\beta' W_{M,h} \right) \right] \right) =$$

$$\sup_{P \in \mathcal{P}^M_h} \left\{ E_P \left[\ell \left(\delta(X, Y) - Kh \right) \right] - \lambda \, \mathrm{KL}(P \| Q_h) \right\},$$

where

$$\mathcal{P}_h^M = \left\{ P \in \Delta(\mathbb{R}^{p+k}) : E_P \left[\prod_{s=1}^{k} \left(Y_s + (\Psi h)_s\right)^{m_s} \right] = E \left[\prod_{s=1}^{k} \xi_s^{m_s} \right] \text{ for all } m \in \mathbb{N}_0^k,\ 1 \leq |m| \leq M \right\}.$$

Since $\mathcal{P}_h^M$ is decreasing in M, the lower bound is bounded below by

$$\inf_{\delta} \sup_{h \in \mathbb{R}^p} \sup_{P \in \mathcal{P}_h^\infty} \left\{ E_P \left[\ell \left(\delta(X,Y) - Kh \right) \right] - \lambda \, \mathrm{KL}(P \| Q_h) \right\},$$

where $\mathcal{P}_h^\infty = \bigcap_{M=1}^\infty \mathcal{P}_h^M$. Because the normal distribution is determined by its moments, $P \in \mathcal{P}_h^\infty$ if and only if $Y + \Psi h \sim N(0, \Omega)$ under P, or equivalently $Y \sim N(-\Psi h, \Omega)$ under P.

Using the chain rule for KL divergence,

$$\mathrm{KL}(P_{X,Y}, Q_{X,Y}) = \mathrm{KL}(P_Y, Q_Y) + E_{P_Y} \left[\mathrm{KL}(P_{X|Y}, Q_{X|Y}) \mid Y \right],$$

we can therefore rewrite the $M = \infty$ problem as

$$\inf_{\delta} \sup_{h \in \mathbb{R}^p} \sup_{P_{X|Y}} E_{Q_{Y,h}} \left[E_{P_{X|Y}} [\ell(\delta(X,Y) - Kh) \mid Y] - \lambda \, \mathrm{KL}(P_{X|Y} \| Q_{X|Y,h}) \mid Y \right].$$

Applying Proposition 4 conditional on Y gives

$$\inf_{\delta} \sup_{h \in \mathbb{R}^p} \lambda \cdot E_{Q_{Y,h}} \left[\log \left(E_{Q_{X|Y,h}} \left[\ell^*(\delta(X,Y) - Kh) \mid Y \right] \right) \right],$$

which is the desired bound. □

B.3 Proofs for Results in Section 4

Proof of Theorem 3 As a first step, it is helpful to note that we can write any distribution $\mathcal{P}_{M,h}$ as the a distribution in $\mathcal{P}_{M,0}$ under a group transformation g, where g takes $Z = (X, Y)$ to $g \circ Z = (X + g, Y - \Psi g)$, and for P the distribution of Z, takes (h, P) to $(h + g, g \circ P)$ for P the distribution of $g \circ Z$.

Lemma 9. $P \in \mathcal{P}_{M,0} \iff g \circ P \in \mathcal{P}_{M,g}$

Proof of Lemma 9 For $W_{m,h} = w_M(Y + \Psi h)$, observe that

$$P \in \mathcal{P}_{M,0} \iff \mathbb{E}_P[w_M(Y)] = 0 \iff \mathbb{E}_{g\circ P}[w_M(Y + \Psi g)] = 0 \iff P \in \mathcal{P}_{M,g},$$

from which the result is immediate. □

This lemma implies that we can parameterize $\mathcal{P}_{M,g}$ by $\mathcal{P}_{M,0}$. Specifically, let $\phi \in \Phi = \mathcal{P}_{M,0}$, and observe that by Lemma 9

$$\sup_h \sup_{P\in\mathcal{P}_{M,h}} E_P\left[\ell\left(\delta\left(X,Y\right) - Kh\right) - \lambda KL(P||Q_h)\right] =$$
$$\sup_{h,\phi} E_{P(h,\phi)}\left[\ell\left(\delta\left(X,Y\right) - Kh\right)\right] - \lambda KL(P(h,\phi)||Q_h),$$

where $P(h,\phi) = h \circ \phi$. Our statistical model is then $\{P(h,\phi) : h \in \mathbb{R}^p, \phi \in \Phi\}$. We can define the group operation on this parameter space by

$$g \circ (h,\phi) = (g + h, \phi),$$

and observe that

$$P(g \circ (h,\phi)) = P(g + h, \phi) = g \circ P(h,\phi),$$

so this definition in the reparametrized model is compatible with our earlier definition.

For the purpose of the decision problem that we're considering, it's without loss of generality to focus on ϕ which has a probability density function with respect to Lebesgue measure. This is because Q_0 is a multivariate normal distribution and if $\phi \not\ll Q_0$, then $KL(P(0,\phi)||Q_0) = +\infty$.

Since the decision problem is invariant, by the generalized Hunt-Stein theorem (Theorem 48.16 of Strasser, 1985) and level-compactness of ℓ, for any δ there exists an equivariant δ^E such that

$$\sup_{g\in\mathbb{R}^p} E_{P(g\circ(h,\phi))}\left[\ell\left(\delta\left(X,Y\right) - K(g+h)\right)\right] \geq E_{P(0,\phi)}\left[\ell\left(\delta^E\left(X,Y\right)\right)\right], \forall\phi$$

Since $g \in G = \mathbb{R}^p$ operates transitively on $H = \mathbb{R}^p$, this is equivalent to

$$\sup_{h\in\mathbb{R}^p} E_{P(h,\phi)}\left[\ell\left(\delta\left(X,Y\right) - Kh\right)\right] \geq E_{P(0,\phi)}\left[\ell\left(\delta^E\left(X,Y\right)\right)\right], \forall\phi$$

Notice that for given ϕ,

$$KL(P(h,\phi)||Q_h) = KL(P(g+h,\phi)||Q_{g+h}).$$

Therefore,

$$\sup_{h\in\mathbb{R}^p} E_{P(h,\phi)}\left[\ell\left(\delta\left(X,Y\right)-Kh\right)\right]-\lambda KL(P(h,\phi)||Q_h)\geq$$
$$E_{P(0,\phi)}\left[\ell\left(\delta^E\left(X,Y\right)\right)\right]-\lambda KL(P(0,\phi)||Q_0),\forall\phi$$

Take supremum over ϕ on both sides of the inequality,

$$\sup_{h\in\mathbb{R}^p,\phi\in\Phi} E_{P(h,\phi)}\left[\ell\left(\delta\left(X,Y\right)-Kh\right)\right]-\lambda KL(P(h,\phi)||Q_h)\geq$$
$$\sup_{\phi\in\Phi} E_{P(0,\phi)}\left[\ell\left(\delta^E\left(X,Y\right)\right)\right]-\lambda KL(P(0,\phi)||Q_0),$$

as we aimed to show. □

Proof of Proposition 7

Part (a). For each β, the minimization of $E_{Q_0}\left[\ell^*(\delta^E(X,Y))\exp(\beta' W_{M,0})\right]$ over $\delta^E\in\mathcal{D}^E$ is equivalent to minimizing $E_{Q_0}\left[\ell^*(\delta^E(X,Y))\frac{\exp(\beta' W_{M,0})}{E_{Q_0}[\exp(\beta' W_{M,0})]}\right]$, which is the problem of finding the best equivariant rule under the tilted distribution with density proportional to $q_0(X,Y)\exp(\beta' W_{M,0})$. This is again a location problem for the group $G=\mathbb{R}^p$, since under the group action

$$(X,Y,h)\mapsto(X+g,Y-\Psi g,h+g),$$

the term $W_{M,h}$ is preserved:

$$W_{M,h+g}(X+g,Y-\Psi g)=W_{M,h}(X,Y).$$

Hence Theorem 6.5 of Eaton (1989) implies that the best equivariant rule is the Bayes rule under the right Haar, that is, flat, prior. The posterior density under this prior is

$$\pi_\beta(h\mid X,Y)=\frac{q_0(X-h,Y+\Psi h)\exp(\beta' W_{M,h})}{\int q_0(X-h',Y+\Psi h')\exp(\beta' W_{M,h'})\,dh'},$$

and the optimal rule is

$$\delta^*_\beta(X,Y) \in \arg\min_{a\in\mathbb{R}^d} \int \ell^*(a - Kh)\pi_\beta(h \mid X, Y)\, dh.$$

Part (b). Let $Z^I = Y + \Psi X$, and observe that this is a maximal invariant for our problem. Since the difference between any two equivariant rules is invariant, and KX is equivariant, any equivariant rule can be written as

$$\delta^E(X,Y) = KX + d(Z^I)$$

for some function $d : \mathbb{R}^k \to \mathbb{R}^d$. Under Q_0, the pair (X, Z^I) is jointly Gaussian with

$$\mathrm{Cov}_{Q_0}(X, Z^I) = \mathrm{Cov}_{Q_0}(X,Y) + \mathrm{Cov}_{Q_0}(X, \Psi X) = -I_0^{-1}\Psi' + I_0^{-1}\Psi' = 0,$$

so X and Z^I are independent. Writing $W_{M,0} = w_M(Y) = w_M(Z^I - \Psi X)$ for the polynomial w_M defining the moment vector, we obtain

$$E_{Q_0}\left[\ell^*(\delta^E(X,Y))\exp(\beta' W_{M,0})\right] = E_{Q_{Z^I,0}}\left[\Gamma_\beta(d(Z^I), Z^I)\right],$$

where

$$\Gamma_\beta(a,z) = E_{Q_{X,0}}\left[\ell^*(KX + a)\exp\left(\beta' w_M(z - \Psi X)\right)\right].$$

For each fixed z, the function $(a,\beta) \mapsto \Gamma_\beta(a,z)$ is jointly lower semicontinuous by Fatou's lemma. Moreover, for every compact set $B \subset \mathbb{R}^b$,

$$\inf_{\beta\in B} \Gamma_\beta(a,z) \to \infty \qquad \text{as } \|a\| \to \infty.$$

To see this, note that $P_{Q_0}\{\|X\| \le 1\} > 0$, that $\ell^*(Kx + a) \to \infty$ uniformly over $\|x\| \le 1$ as $\|a\| \to \infty$, and that

$$\inf_{\beta\in B,\ \|x\|\le 1} \exp\left(\beta' w_M(z - \Psi x)\right) > 0$$

because B and $\{z - \Psi x : \|x\| \le 1\}$ are compact. Consequently, we may define

$$g_\beta(z) = \inf_{a\in\mathbb{R}^d} \Gamma_\beta(a,z) = \min_{a\in\mathbb{R}^d} \Gamma_\beta(a,z),$$

which is lower semicontinious in (β, z) by Berge's theorem of the maximum, and

$$m(\beta) = \min_{\delta^E \in \mathcal{D}^E} E_{Q_0}\left[\ell^*(\delta^E(X,Y))\exp(\beta' W_{M,0})\right] = E_{Q_{Z^I,0}}\left[g_\beta(Z^I)\right].$$

Since $g_\beta(Z^I) \geq 0$, lower semicontinuity of $g_\beta(z)$ and Fatou's lemma imply that m is lower semicontinuous. In addition, since $\ell^* \geq 1$,

$$m(\beta) \geq E_{Q_0}\left[\exp(\beta' W_{M,0})\right].$$

By Lemma 7 with $A = 1$, the right-hand side has compact sublevel sets. Hence m also has compact sublevel sets. Therefore m attains its infimum at some $\beta^* \in \mathbb{R}^b$.

Part (c). By part (a), $\delta^*_{\beta^*}$ minimizes the joint objective (14) over (δ^E, β). By Theorem 3, the minimax value over all decision rules equals the minimax value over equivariant rules, and by the duality of Proposition 5 the latter equals (14). Hence $\delta^*_{\beta^*}$ is minimax optimal. □

Proof of Proposition 8

Part (a). The function ℓ is convex by Assumption 2, so $\Gamma \mapsto \ell(\Gamma\phi)$ is convex in Γ. The product $\ell^*(\Gamma\phi)\exp(\beta' W_{M,0}) = \exp(\ell(\Gamma\phi)/\lambda + \beta' W_{M,0})$ is the exponential of a function that is jointly convex in (Γ, β), and the exponential of a convex function is convex. Taking expectations preserves convexity.

Part (b). Fix $\Gamma_1, \Gamma_2 \in \mathbb{R}^{d\times J}$ and $t \in [0,1]$, and let $\Gamma_t = t\Gamma_1 + (1-t)\Gamma_2$. Since $\Gamma \mapsto \Gamma\phi(X,Y)$ is linear and ℓ is convex, $\ell(\Gamma_t\phi) \leq t\,\ell(\Gamma_1\phi) + (1-t)\,\ell(\Gamma_2\phi)$ pointwise. Exponentiating and using $\ell^*(u) = \exp(\ell(u)/\lambda)$ gives

$$\ell^*(\Gamma_t\phi) \leq \ell^*(\Gamma_1\phi)^t\,\ell^*(\Gamma_2\phi)^{1-t}$$

pointwise. Hence, by Hölder's inequality,

$$E_{Q_{X|Y,0}}[\ell^*(\Gamma_t\phi) \mid Y] \leq E_{Q_{X|Y,0}}[\ell^*(\Gamma_1\phi) \mid Y]^t E_{Q_{X|Y,0}}[\ell^*(\Gamma_2\phi) \mid Y]^{1-t}.$$

Taking logs yields

$$\log E_{Q_{X|Y,0}}[\ell^*(\Gamma_t\phi) \mid Y] \leq t \log E_{Q_{X|Y,0}}[\ell^*(\Gamma_1\phi) \mid Y] + (1-t)\log E_{Q_{X|Y,0}}[\ell^*(\Gamma_2\phi) \mid Y].$$

Thus $\Gamma \mapsto \log E_{Q_{X|Y,0}}[\ell^*(\Gamma\phi) \mid Y]$ is convex for each Y, and taking expectation over $Q_{Y,0}$ preserves convexity. □

Proof of Proposition 9 We interpret the objective throughout in the extended-real sense. If every equivariant rule has infinite risk, the proposition is immediate. Hence we may restrict attention to the case where at least one equivariant rule has finite risk.

Let $Z^I = Y + \Psi X$, and observe that this is a maximal invariant for our problem. Since the difference between any two equivariant rules is invariant, and KX is equivariant, any equivariant rule can be written as

$$\delta^E(X, Y) = KX + d(Z^I)$$

for some function $d : \mathbb{R}^k \to \mathbb{R}^d$.

For $M < \infty$, define

$$\mathcal{R}_M(d, \beta) = E_{Q_0}\left[\exp\left(\frac{1}{\lambda}\|KX + d(Z^I)\|^2 + \beta' W_{M,0}\right)\right].$$

Part (a): $M = 0$ and $M = 1$ For $M = 0$, there is no β. Since $X \perp Z^I$ under Q_0, we have

$$\mathcal{R}_0(d) = E_{Q_{Z^I,0}}\left[\varphi(d(Z^I))\right], \text{ where } \varphi(a) = E_{Q_{X,0}}\left[\exp\left(\frac{1}{\lambda}\|KX + a\|^2\right)\right].$$

The map $a \mapsto \varphi(a)$ is strictly convex, and by symmetry of the distribution of X under $h = 0$,

$$\nabla\varphi(0) = \frac{2}{\lambda} E_{Q_{X,0}}\left[KX \exp\left(\frac{1}{\lambda}\|KX\|^2\right)\right] = 0.$$

Hence φ is uniquely minimized at $a = 0$, so $\mathcal{R}_0(d)$ is minimized by $d \equiv 0$, i.e. $\delta^*(X, Y) = KX$.

For $M = 1$, write $W_{1,0} = Y$. Using $Y = Z^I - \Psi X$ and $X \perp Z^I$:

$$\mathcal{R}_1(d, \beta) = E_{Q_{Z^I,0}}\left[e^{\beta' Z^I} H_\beta(d(Z^I))\right], \qquad H_\beta(a) = E_{Q_{X,0}}\left[\exp\left(\frac{1}{\lambda}\|KX + a\|^2 - \beta'\Psi X\right)\right].$$

For each β, $H_b(\cdot)$ has a unique minimizer a_b so long as the minimized value is finite, and

$$\inf_d \mathcal{R}_1(d, \beta) = E_{Q_0}\left[e^{\beta' Z^I}\right] H_b(a_b).$$

Thus

$$\inf_{d,\beta} \mathcal{R}_1(d,\beta) = \inf_{a,\beta} E_{Q_0}\left[e^{\beta' Z^I}\right] H_\beta(a).$$

The objective is jointly convex in (a,β). At $(a,\beta)=(0,0)$,

$$\nabla_a \mathcal{R}_1(0,0) = \frac{2}{\lambda} E_{Q_0}\left[KX \exp\left(\frac{1}{\lambda}\|KX\|^2\right)\right] = 0,$$

and

$$\nabla_\beta \mathcal{R}_1(0,0) = E_{Q_0}\left[Y \exp\left(\frac{1}{\lambda}\|KX\|^2\right)\right] = 0,$$

again by symmetry of the distribution of X. Convexity therefore implies $(a,b)=(0,0)$ is globally optimal, so $\delta^*(X,Y) = KX$ for $M=1$ as well.

Part (b): $M \geq 2$ Define the primal constrained multiplier risk

$$\mathcal{R}_M(\delta^E) = \sup_{P \in \mathcal{P}_{M,0}} \left\{ E_P\left[\|\delta^E(X,Y)\|^2\right] - \lambda KL(P\|Q_0)\right\}.$$

For the $M=\infty$ problem, write

$$R_\infty(d) = \lambda E_{Q_{Y,0}}\left[\log G_d(Y)\right], \qquad G_d(y) = E_{Q_{X|Y,0}}\left[\exp\left(\frac{1}{\lambda}\|KX + d(Z^I)\|^2\right) \,\middle|\, Y = y\right],$$

where $\delta^E(X,Y) = KX + d(Z^I)$.

Lemma 10. *Suppose $\ell(u) = \|u\|^2$ and $R_\infty(d) < \infty$. Define*

$$w_d(X,Y) = \frac{\exp\left(\frac{1}{\lambda}\|KX + d(Z^I)\|^2\right)}{G_d(Y)}$$

and

$$g_d(Z^I) = 2E_{Q_0}\left[\left(KX + d(Z^I)\right) w_d(X,Y) \mid Z^I\right].$$

Then for every measurable perturbation $\gamma(Z^I)$ such that $R_\infty(d+\gamma) < \infty$ and $E_{Q_0}\left[\left|g_d(Z^I)'\gamma(Z^I)\right|\right] < \infty$,

$$R_\infty(d+\gamma) \geq R_\infty(d) + E_{Q_0}\left[g_d(Z^I)'\gamma(Z^I)\right].$$

In particular, if $g_d(Z^I) = 0$ almost surely, then d is globally optimal for the $M=\infty$ problem.

Proof of Lemma 10 For each y, define the convex functional

$$\mathcal{F}_y(u) = \lambda \log E_{Q_{X|Y,0}} \left[\exp\left(\frac{1}{\lambda} u(X, y) \right) \mid Y = y \right]$$

on the set of $u(\cdot, y)$ for which the expectaion is finite. Let $u_d(X, Y) = \|KX + d(Z^I)\|^2$ and define the tilted conditional density

$$w_d(X, Y) = \frac{\exp\left(\frac{1}{\lambda} u_d(X, Y)\right)}{E_{Q_{X|Y,0}}\left[\exp\left(\frac{1}{\lambda} u_d(X, Y)\right) \mid Y\right]}.$$

Then for any perturbation $v(X, Y)$ such that both sides are finite,

$$\mathcal{F}_Y(u_d + v) - \mathcal{F}_Y(u_d) = \lambda \log E_{Q_{X|Y,0}} \left[\exp\left(\frac{1}{\lambda} v(X, Y) \right) w_d(X, Y) \mid Y \right].$$

If we apply Jensen's inequality to the tilted measure with Radon-Nikodym derivative $w_d(\cdot, Y)$ with respect to $Q_{X|Y,0}(\cdot \mid Y)$, we thus have

$$\mathcal{F}_Y(u_d + v) - \mathcal{F}_Y(u_d) \geq E_{Q_{X|Y,0}}[v(X, Y) w_d(X, Y) \mid Y].$$

Now take

$$v(X, Y) = u_{d+\gamma}(X, Y) - u_d(X, Y) = 2\left(KX + d(Z^I)\right)' \gamma(Z^I) + \|\gamma(Z^I)\|^2.$$

Taking expectations over $Q_{Y,0}$ yields $R_\infty(d + \gamma) - R_\infty(d) \geq E_{Q_0}\left[v(X, Y) w_d(X, Y)\right]$.

Define

$$g_d(Z^I) = 2E_{Q_0}\left[\left(KX + d(Z^I)\right) w_d(X, Y) \mid Z^I\right].$$

Then

$$E_{Q_0}[v(X, Y) w_d(X, Y)] = E_{Q_0}\left[g_d(Z^I)'\gamma(Z^I)\right] + E_{Q_0}\left[w_d(X, Y)\|\gamma(Z^I)\|^2\right] \geq E_{Q_0}\left[g_d(Z^I)'\gamma(Z^I)\right].$$

Therefore

$$R_\infty(d + \gamma) \geq R_\infty(d) + E_{Q_0}\left[g_d(Z^I)'\gamma(Z^I)\right].$$

Thus g_d is a subgradient of R_∞ at d. If $g_d(Z^I) = 0$ almost surely, then 0 is a subgradient at d, so convexity of R_∞ implies that d is globally optimal. □

Now consider the linear class

$$\delta_C(X,Y) = KX + CZ^I, \qquad C \in \mathbb{R}^{d\times k},$$

and define $r(C) = R_\infty(\delta_C)$. Since $\lambda \log E[e^{U/\lambda} \mid Y] \geq E[U \mid Y]$ by Jensen's inequality,

$$r(C) \geq E_{Q_0}\left[\|KX + CZ^I\|^2\right] = E_{Q_0}\left[\|KX\|^2\right] + \mathrm{Trace}\left(C\ \mathrm{Var}_{Q_0}(Z^I)\, C'\right),$$

where we used $E_{Q_0}[XZ^{I\prime}] = 0$. Let $\Sigma_I = \mathrm{Var}_{Q_0}(Z^I)$ and let Π be the orthogonal projection onto the range of Σ_I. Since $Z^I = \Pi Z^I$ almost surely, $r(C) = r(C\Pi)$, and on this reduced space Σ_I is positive definite. Hence $\mathrm{Trace}(C\Sigma_I C') \to \infty$ whenever $\|C\Pi\| \to \infty$, so sublevel sets of r are compact, and there exists a minimizing value C^* whenever $\inf_C r(C)$ is finite.

Define

$$\frac{dP^\infty_C}{dQ_0}(X,Y) = \frac{\exp\left(\frac{1}{\lambda}\|\delta_C(X,Y)\|^2\right)}{E_{Q_{X|Y,0}}\left[\exp\left(\frac{1}{\lambda}\|\delta_C(X,Y)\|^2\right) \mid Y\right]}.$$

For any matrix $D \in \mathbb{R}^{d\times k}$, consider $\gamma_D(Z^I) = DZ^I$. By Lemma 10, optimality of C^* relative to the class of linear rules implies

$$0 = E_{Q_0}\left[g_{d_{C^*}}(Z^I)' DZ^I\right] = 2E_{P^\infty_{C^*}}\left[\delta_{C^*}(X,Y)' DZ^I\right] \quad \text{for all } D,$$

and thus

$$E_{P^\infty_{C^*}}\left[\delta_{C^*}(X,Y) Z^{I\prime}\right] = 0.$$

Since δ_C is linear and we consider quadratic loss, (X,Y) is Gaussian under $P^\infty_{C^*}$, so $(\delta_{C^*}(X,Y), Z^I)$ is jointly Gaussian. Moreover, $E_{P^\infty_{C^*}}[\delta_{C^*}(X,Y)] = 0$, since the marginal distribution of Y under $P^\infty_{C^*}$ is $Q_{Y,0}$ and the tilted conditional mean of X given Y is linear in Y with no constant term. Since uncorrelated jointly Gaussian random vectors are independent,

$$E_{P^\infty_{C^*}}\left[\delta_{C^*}(X,Y) \mid Z^I\right] = 0.$$

Equivalently, $g_{d_{C^*}}(Z^I) = 0$ almost surely, so Lemma 10 implies that δ_{C^*} is globally optimal for the $M = \infty$ problem over all equivariant rules.

Next fix any linear rule δ_C and define

$$\phi_C(y) = \lambda \log E_{Q_{X|Y,0}}\left[\exp\left(\frac{1}{\lambda}\|\delta_C(X,Y)\|^2\right) \mid Y = y\right].$$

Since $X|Y = y$ is Gaussian and δ_C is affine in X, $\phi_C(y)$ corresponds to the moment generating function of a non-central χ^2 distribution and (when finite) is quadratic in y.

For any $P \in \mathcal{P}_0^2$, decompose $P = P_{X|Y}P_Y$ and $Q_0 = Q_{X|Y,0}Q_{Y,0}$. Then

$$\begin{aligned}&E_P\left[\|\delta_C(X,Y)\|^2\right] - \lambda\,\mathrm{KL}(P\|Q_0)\\ &= E_{P_Y}\left[E_{P_{X|Y}}\left[\|\delta_C(X,Y)\|^2 \mid Y\right] - \lambda\,\mathrm{KL}\left(P_{X|Y}\|Q_{X|Y,0}(\cdot \mid Y)\right)\right] - \lambda\,\mathrm{KL}(P_Y\|Q_{Y,0}).\end{aligned}$$

Applying Proposition 4 conditional on Y gives

$$E_{P_{X|Y}}\left[\|\delta_C(X,Y)\|^2 \mid Y\right] - \lambda\,\mathrm{KL}\left(P_{X|Y}\|Q_{X|Y,0}(\cdot \mid Y)\right) \le \phi_C(Y),$$

from which it follows that

$$E_P\left[\|\delta_C(X,Y)\|^2\right] - \lambda\,\mathrm{KL}(P\|Q_0) \le E_{P_Y}[\phi_C(Y)] - \lambda\,\mathrm{KL}(P_Y\|Q_{Y,0}).$$

Because $P \in \mathcal{P}_0^2$ imposes $E_P[Y] = 0$ and $E_P[YY'] = \Omega$, and ϕ_C is quadratic,

$$E_{P_Y}[\phi_C(Y)] = E_{Q_{Y,0}}[\phi_C(Y)].$$

Hence

$$\mathcal{R}_2(\delta_C) \le E_{Q_{Y,0}}[\phi_C(Y)] = \mathcal{R}_\infty(\delta_C).$$

The reverse inequality $\mathcal{R}_2(\delta_C) \ge \mathcal{R}_\infty(\delta_C)$ is immediate from $\mathcal{P}_0^\infty \subseteq \mathcal{P}_0^2$, where $\mathcal{P}_0^\infty = \cap_{m\ge1}\mathcal{P}_0^m$. Therefore $\mathcal{R}_2(\delta_C) = \mathcal{R}_\infty(\delta_C)$. Moreover, P_C^∞ attains $\mathcal{R}_\infty(\delta_C)$, is Gaussian, and has $Y \sim Q_{Y,0}$, so $P_C^\infty \in \mathcal{P}_0^M$ for every $M \ge 2$.

Using $\mathcal{P}_0^\infty \subseteq \mathcal{P}_0^M \subseteq \mathcal{P}_0^2$ for $M \ge 2$, we obtain for every linear rule δ_C,

$$\mathcal{R}_M(\delta_C) = \mathcal{R}_\infty(\delta_C).$$

Applying this at $C = C^*$,

$$\mathcal{R}_M(\delta_{C^*}) = \mathcal{R}_\infty(\delta_{C^*}) = \inf_C \mathcal{R}_\infty(\delta_C).$$

Finally, for any equivariant δ^E and $M \ge 2$,

$$\mathcal{R}_M(\delta^E) \ge \mathcal{R}_\infty(\delta^E) \ge \mathcal{R}_\infty(\delta_{C^*}) = \mathcal{R}_M(\delta_{C^*}),$$

so δ_{C^*} is minimax optimal for each $M \geq 2$, and the optimal value is the same for all $M \geq 2$. □

Proof of Proposition 10 For $h \in \mathbb{R}^p$, define $T_{n,h} = \sqrt{n}\left(\delta_n^c - \kappa(\theta_{n,h})\right)$. By the definition of δ_n^c,

$$T_{n,h} = \sqrt{n}\left(\kappa(\hat{\theta}_n^{MLE}) - \kappa(\theta_{n,h})\right) + \delta^c\left(0, \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(\hat{\theta}_n^{MLE}, X_i); \hat{\Sigma}_n\right).$$

By differentiability of κ at θ_0 and Assumption 3,

$$\sqrt{n}\left(\kappa(\hat{\theta}_n^{MLE}) - \kappa(\theta_{n,h})\right) \overset{Q_{n,h}}{\underset{d}{\to}} K(X - h).$$

Again by Assumption 3, under $Q_{n,h}$,

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(\hat{\theta}_n^{MLE}, X_i), \hat{\Sigma}_n\right) \overset{Q_{n,h}}{\underset{d}{\to}} (Y + \Psi X, \Sigma),$$

so continuity of δ^c and the continuous mapping theorem give

$$\delta^c\left(0, \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(\hat{\theta}_n^{MLE}, X_i); \hat{\Sigma}_n\right) \overset{Q_{n,h}}{\underset{d}{\to}} \delta^c(0, Y + \Psi X; \Sigma).$$

Using equivariance with $g = -X$,

$$\delta^c(0, Y + \Psi X; \Sigma) = \delta^c(X, Y; \Sigma) - KX.$$

Therefore

$$T_{n,h} \overset{Q_{n,h}}{\underset{d}{\to}} K(X - h) + \delta^c(0, Y + \Psi X; \Sigma) = \delta^c(X, Y; \Sigma) - Kh.$$

as claimed. □

Proof of Corollary 2 For $h \in \mathbb{R}^p$, let

$$T_{n,h} = \sqrt{n}\left(\delta_n^c - \kappa(\theta_{n,h})\right), \quad T_h = \delta^c(X, Y; \Sigma) - Kh,$$

and define $f(t, w; \beta) = \ell^*(t)\exp(\beta' w)$. By Proposition 10 and the definition of $W_{M,n,h}$, under $Q_{n,h}$,

$$(T_{n,h}, W_{M,n,h}) \overset{Q_{n,h}}{\underset{d}{\to}} (T_h, W_{M,h}).$$

By Lemma 8, for each h,

$$\liminf_{n\to\infty} \inf_{\beta} E_{Q_{n,h}}\left[f(T_{n,h}, W_{M,n,h}; \beta)\right] \geq \inf_{\beta} E_{Q_h}\left[f(T_h, W_{M,h}; \beta)\right]. \tag{27}$$

Next, by equivariance of δ^c and the Gaussian shift structure of the limit experiment,

$$E_{Q_h}\left[f(T_h, W_{M,h}; \beta)\right] = E_{Q_0}\left[\ell^*\left(\delta^c(X, Y; \Sigma)\right) \exp\left(\beta' W_{M,0}\right)\right]$$

for every h and β. Hence $\beta^{*,c}$ from Assumption 4 minimizes the right-hand side for every h. Therefore, for each h,

$$\begin{aligned}\limsup_{n\to\infty} \inf_{\beta} E_{Q_{n,h}}\left[f(T_{n,h}, W_{M,n,h}; \beta)\right] &\leq \lim_{n\to\infty} E_{Q_{n,h}}\left[f(T_{n,h}, W_{M,n,h}; \beta^{*,c})\right] \\ &= E_{Q_h}\left[f(T_h, W_{M,h}; \beta^{*,c})\right] = \inf_{\beta} E_{Q_h}\left[f(T_h, W_{M,h}; \beta)\right], \end{aligned} \tag{28}$$

where the equality in the middle follows from Assumption 4. Combining (27) and (28), for each h,

$$\lim_{n\to\infty} \inf_{\beta} E_{Q_{n,h}}\left[f(T_{n,h}, W_{M,n,h}; \beta)\right] = \inf_{\beta} E_{Q_h}\left[f(T_h, W_{M,h}; \beta)\right].$$

Since I is finite, taking $\sup_{h\in I}$ preserves convergence, and applying $\lambda \log(\cdot)$ gives

$$\lim_{n\to\infty} \sup_{h\in I} \inf_{\beta} \lambda \cdot \log\left(E_{Q_{n,h}}\left[\ell^*(T_{n,h}) \exp\left(\beta' W_{M,n,h}\right)\right]\right) = \sup_{h\in I} \inf_{\beta} \lambda \cdot \log\left(E_{Q_h}\left[\ell^*(T_h) \exp\left(\beta' W_{M,h}\right)\right]\right).$$

This is the first claim.

For the asymptotic optimality claim, choose δ^c optimal in the limit experiment. The convergence argument above gives that the asymptotic risk is

$$\sup_{h} \inf_{\beta} \lambda \cdot \log\left(E_{Q_h}\left[\ell^*\left(\delta^c(X, Y) - Kh\right) \exp\left(\beta' W_{M,h}\right)\right]\right).$$

By Theorem 2, however, this is the best attainable risk, proving the claim. □