Electronic Companion to *Risk Guarantees for End-to-End Prediction and Optimization Processes*

EC.1. A Note on the Regularity of X^* and x^*

We denote the power set, the collection of all subsets of X, as 2^X. An important property that we exploit is that the argmin mapping $X^*(d)$ is, in a sense, well-behaved as we change d. More precisely, the sense of regularity that we use is upper semicontinuity, which stems from a result in perturbation analysis (Bonnans and Shapiro 2000).

DEFINITION EC.1. A multivalued function $F : \mathbb{R}^m \to 2^X$ is *upper semi-continuous* at a point $d \in \mathbb{R}^m$ if, for any open set U containing $F(d)$, there exists an open set U_d containing d such that for all $d' \in U_d$, $F(d') \subseteq U$. Equivalently, F is upper semi-continuous if, for any closed set V, the following set is closed:

$$\{d \in \mathbb{R}^m : F(d) \cap V \neq \emptyset\}.$$

LEMMA EC.1. *Suppose X is compact. Then the multivalued mapping $X^* : \mathbb{R}^m \to 2^X$ is upper semi-continuous.*

Proof. This follows immediately from verifying the conditions of Bonnans and Shapiro (2000, Proposition 4.4), which are straightforward to check due to the fact that the domain X does not change with the vector d. □

We can use Lemma EC.1 to show the existence of a measurable selection $x^*(d) \in X^*(d)$ via an application of the Kuratowski–Ryll-Nardzewski theorem on the existence of measurable selectors for multivalued mappings. We use the version stated in Bogachev (2007, Theorem 6.9.3).

LEMMA EC.2. *Suppose X is compact. Then there exists a measurable mapping $x^* : \mathbb{R}^m \to X$ such that $x^*(d) \in X^*(d)$ for all $d \in \mathbb{R}^m$.*

Proof. Consider the multivalued function $X^* : \mathbb{R}^m \to 2^X$ defined by $X^*(d) = \arg\min_{x \in X} d^\top x$. Note that since $d^\top x$ is continuous, $X^*(d) = \{x \in X : d^\top x = \min_{x' \in X} d^\top x'\}$ is closed (it is the inverse of a singleton). Now consider an open set U, and the sets

$$\hat{X}^*(U) := \{d \in \mathbb{R}^m : X^*(d) \cap U \neq \emptyset\}.$$

It is known that U can be represented as the countable union of closed sets: $U = \bigcup_{k \in \mathbb{N}} V_k$ where V_k are closed. Thus, we can write

$$\hat{X}^*(U) = \{d \in \mathbb{R}^m : \exists k \in \mathbb{N} \text{ s.t. } X^*(d) \cap V_k \neq \emptyset\} = \bigcup_{k \in \mathbb{N}} \{d \in \mathbb{R}^m : X^*(d) \cap V_k \neq \emptyset\}.$$

Now, since $X^*(d)$ is upper semicontinuous, $\{d \in \mathbb{R}^m : X^*(d) \cap U_k \neq \emptyset\}$ is closed, hence $\hat{X}^*(U)$ is a countable union of closed sets, hence measurable. This shows that $X^*(\cdot)$ satisfies the conditions of Bogachev (2007, Theorem 6.9.3), therefore there exists a measurable selection $x^*(d) \in X^*(d)$ for all $d \in \mathbb{R}^m$. □

Furthermore, we can show that *any* selection x^* is at least *Lebesgue* measurable, using the following result of Drusvyatskiy and Lewis (2011).

LEMMA EC.3 (**Drusvyatskiy and Lewis (2011, Corollary 3.5)**). *The set*

$$D := \{d \in \mathbb{R}^m : X^*(d) \text{ is not a singleton}\}$$

has Lebesgue measure zero.

LEMMA EC.4. *Any selection $x^* : \mathbb{R}^m \to X$ such that $x^*(d) \in X^*(d)$ for all $d \in \mathbb{R}^m$ is Lebesgue measurable.*

Proof. Lemma EC.2 tells us that there exists one such measurable selection $\bar{x}^*$. Consider another selection x^*. Then by Lemma EC.3, $\bar{x}^*$ and x^* differ on at most a set D with Lebesgue measure 0, which is Lebesgue measurable. Furthermore, all subsets of D are also Lebesgue measurable, so x^* must be Lebesgue measurable. □

In order for our expectations to be well-defined, we make the following assumption.

ASSUMPTION EC.1. *Any probability distribution $\mathbb{P}$ is defined on the σ-algebra of Lebesgue measurable sets.*

This is not practically restrictive, since any probability distribution we encounter in practice can be written as a mixture of a distribution which is absolutely continuous with respect to Lebesgue measure (i.e., it has a density function), and a discrete distribution supported on a countable set. Such a probability distribution is Lebesgue measurable.

EC.2. Proof of Results from Section 3

Proof of Lemma 1. The measurability of $w \mapsto \mathbb{E}[c \mid w]$ is obvious by definition of the conditional expectation. Fix some measurable $g : W \to \mathbb{R}^m$. Observe that for $w \in W$,

$$\begin{aligned}
\mathbb{E}[L(g(w), c) \mid w] &= \mathbb{E}\left[f(x^*(g(w))) + c^\top x^*(g(w)) - \min_{x \in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&= \mathbb{E}\left[f(x^*(g(w))) \mid w\right] + \mathbb{E}\left[c \mid w\right]^\top x^*(g(w)) - \mathbb{E}\left[\min_{x \in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&= \mathbb{E}\left[f(x^*(g(w))) \mid w\right] + g^*(w)^\top x^*(g(w)) - \mathbb{E}\left[\min_{x \in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&\geq f(x^*(g^*(w))) + g^*(w)^\top x^*(g^*(w)) - \mathbb{E}\left[\min_{x \in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&= f(x^*(g^*(w))) + \mathbb{E}\left[c \mid w\right]^\top x^*(g^*(w)) - \mathbb{E}\left[\min_{x \in X}\left\{f(x) + c^\top x\right\} \mid w\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[f(x^*(g^*(w))) + c^\top x^*(g^*(w)) - \min_{x\in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&= \mathbb{E}[L(g^*(w), c) \mid w],
\end{aligned}
$$

where the inequality follows from the definition of $x^*(\cdot)$. Integrating both sides of this relation over $w \in W$ gives $R(g, \mathbb{P}) \geq R(g^*, \mathbb{P})$. Thus, g^* is the minimizer of $R(g, \mathbb{P})$.

The second result follows because

$$
\begin{aligned}
\min_{d'\in\mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] &= \min_{d'\in\mathbb{R}^m} \left\{f(x^*(d')) + \mathbb{E}[c \mid w]^\top x^*(d')\right\} - \mathbb{E}\left[\min_{x\in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&= \min_{d'\in\mathbb{R}^m} \left\{f(x^*(d')) + g^*(w)^\top x^*(d')\right\} - \mathbb{E}\left[\min_{x\in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&= f(x^*(g^*(w))) + g^*(w)^\top x^*(g^*(w)) - \mathbb{E}\left[\min_{x\in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&= f(x^*(g^*(w))) + \mathbb{E}\left[c \mid w\right]^\top x^*(g^*(w)) - \mathbb{E}\left[\min_{x\in X}\left\{f(x) + c^\top x\right\} \mid w\right] \\
&= \mathbb{E}\left[L(g^*(w), c) \mid w\right],
\end{aligned}
$$

and then integrating both sides over $w \in W$ gives $\mathbb{E}\left[\min_{d'\in\mathbb{R}^m} \mathbb{E}[L(d', c) \mid w]\right] = \mathbb{E}[L(g^*(w), c)] = R(\mathbb{P})$. □

EC.3. Proof of Theorem 1

Define

$$
\delta_\ell(\epsilon, w; \mathbb{P}) := \inf_{d\in\mathbb{R}^m} \left\{\mathbb{E}[\ell(d, c) \mid w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] : \mathbb{E}[L(d, c) \mid w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] \;\geq\; \epsilon\right\}. \tag{EC.1}
$$

Note that if ℓ is $\mathbb{P}$-calibrated, then $\delta_\ell(\epsilon, w; \mathbb{P}) > 0$ for all $\epsilon > 0, w \in W$ by taking the contrapositive of the implication in Definition 3. In order to prove Theorem 1, we first verify measurability for δ_ℓ.

Lemma EC.5. *Suppose ℓ is measurable and satisfies Assumption 1, and that X is compact. For any $\epsilon > 0$, the function $\delta_\ell(\epsilon, \cdot; \mathbb{P}) : W \to \mathbb{R}$ is measurable.*

Proof. Consider the set

$$
W_r := \{w \in W : \delta_\ell(\epsilon, w; \mathbb{P}) \leq r\}.
$$

Showing measurability of $\delta_\ell(\epsilon, \cdot; \mathbb{P})$ boils down to showing that W_r is measurable. Rewrite

$$
\begin{aligned}
W_r &= \left\{w \in W : \forall k \in \mathbb{N},\ \exists d \in \mathbb{R}^m \text{ s.t. } \begin{array}{r} \mathbb{E}[\ell(d, c) \mid w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] \leq r + 1/k \\ \mathbb{E}[L(d, c) \mid w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] \;\geq\; \epsilon \end{array}\right\} \\
&= \bigcap_{k\in\mathbb{N}} \left\{w \in W : \exists d \in \mathbb{R}^m \text{ s.t. } \begin{array}{r} \mathbb{E}[\ell(d, c) \mid w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] \leq r + 1/k \\ \mathbb{E}[L(d, c) \mid w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] \;\geq\; \epsilon \end{array}\right\}
\end{aligned}
$$

To this end, first consider the subset

$$W_L(\epsilon) = \left\{(w,d) \in W \times \mathbb{R}^m : \mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] \ \geq \ \epsilon \right\}$$
$$= \left\{(w,d) \in W \times \mathbb{R}^m : f(x^*(d)) + \mathbb{E}[c \mid w]^\top x^*(d) - \min_{x \in X} \left\{f(x) + \mathbb{E}[c \mid w]^\top x\right\} \ \geq \ \epsilon \right\}.$$

This is measurable since $\mathbb{E}[c \mid w]$ is measurable in w by definition of conditional expectation, f is continuous hence measurable, and we have assumed $x^*(d)$ is measurable in d, which is possible by Lemma EC.2.

Now consider the subset

$$W_\ell(\alpha) = \left\{(w,d) \in W \times \mathbb{R}^m : \mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] \leq \alpha \right\}.$$

First observe that the function h defined by $h(w,d) = \mathbb{E}[\ell(d,c) \mid w]$ is continuous in d and measurable in w. Continuity in d follows because $\ell(d,c)$ is convex in d, and $h(w,d)$ is finite for any w by Assumption 1, and all convex functions are continuous in the relative interiors of their domains (see e.g., Rockafellar (1970, Theorem 10.1)). Measurability follows from measurability of ℓ and the definition of conditional expectation.

We now show that h is jointly measurable in (w,d) by showing that it is a pointwise limit of measurable functions. For $k \in \mathbb{N}$, consider the box $B_k := [-k,k]^m \subset \mathbb{R}^m$ and a finite set of grid points $G_k \subset B_k$ such that any point $d \in B_k$ is at most distance $1/k$ away from a grid point in Euclidean norm. If $d \in B_k$, define $h_k(w,d) = h(w,g)$ where $g \in B_k$ is the closest grid point to d (with ties broken arbitrarily), and if $d \notin B_k$ define $h_k(w,d) = 0$. Note that fixing g, $w \mapsto h_g(w) := h(w,g)$ is measurable in w. Now, h_k is the sum of finitely many functions of the form $\mathbf{1}_D(d)h_g(w)$ for some measurable set D and grid point g. It is easy to check that this is measurable, therefore h_k is measurable. Furthermore, by continuity of h in d, $h_k(w,d) \to h(w,d)$ pointwise. Therefore, h is measurable. Finally, the function $(w,d) \mapsto \min_{d' \in \mathbb{R}^m} h(w,d')$ is measurable because by continuity of h in d, we can write

$$\left\{(w,d) : \min_{d' \in \mathbb{R}^m} h(w,d') \leq \alpha \right\} = \bigcup_{d' \in D_*, k \in \mathbb{N}} \{(w,d) : h(w,d') \leq \alpha + 1/k\}$$

where D_* is a countable dense subset of $\mathbb{R}^m$ (e.g., $\mathbb{Q}^m$). This shows that $W_\ell(\alpha)$ is measurable because the function $h(w,d) - \min_{d' \in \mathbb{R}^m} h(w,d') = \mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w]$ is measurable.

Now notice that the set

$$\left\{(w,d) \in W \times \mathbb{R}^m : \begin{array}{c} \mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] \leq r + 1/k \\ \mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] \ \geq \ \epsilon \end{array} \right\} = W_\ell(r+1/k) \cap W_L(\epsilon)$$

is measurable. Therefore, its projection onto W is measurable, which is

$$\left\{ w \in W : \exists d \in \mathbb{R}^m \text{ s.t. } \begin{array}{r} \mathbb{E}[\ell(d,c) \mid w] - \min\limits_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] \leq r + 1/k \\ \mathbb{E}[L(d,c) \mid w] - \min\limits_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] \ \geq \ \epsilon \end{array} \right\}.$$

This shows that W_r is measurable, concluding our proof. □

Proof of Theorem 1. We wish to apply a result of Steinwart (2007, Theorem 2.8), for which we need to show that there exists measurable functions $b : W \to \mathbb{R}$ and and $\delta : (0,\infty) \times W \to (0,\infty)$ such that $\mathbb{E}[|b(w)|] < \infty$, for any $w \in W$

$$\mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] \leq b(w),$$

and for any $\epsilon > 0$ and predictor $g : W \to \mathbb{R}^m$,

$$\begin{aligned} &\left\{ w \in W : \mathbb{E}[\ell(g(w),c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] < \delta(\epsilon, w) \right\} \\ &\subseteq \left\{ w \in W : \mathbb{E}[L(g(w),c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] < \epsilon \right\}. \end{aligned}$$

We first find b. Let Ω be the ℓ_∞-diameter of the set X, which is finite since X is compact. Observe that for any $d \in \mathbb{R}^m$,

$$\begin{aligned} \mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] &= f(x^*(d)) + \mathbb{E}[c \mid w]^\top x^*(d) - \min_{x \in X} \left\{ f(x) + \mathbb{E}[c \mid w]^\top x \right\} \\ &\leq \max_{x, x' \in X} \left\{ f(x) - f(x') + \mathbb{E}[c \mid w]^\top (x' - x) \right\} \\ &\leq \max_{x, x' \in X} \left\{ f(x) - f(x') + \|\mathbb{E}[c \mid w]\|_1 \|x' - x\|_\infty \right\} \\ &\leq \Omega \|\mathbb{E}[c \mid w]\|_1 + \max_{x, x' \in X} \left\{ f(x) - f(x') \right\}. \end{aligned}$$

Therefore, we can define $b(w) := \Omega \|\mathbb{E}[c \mid w]\|_1 + \max_{x,x' \in X} \{f(x) - f(x')\}$ for each $w \in W$, which is integrable as $\mathbb{E}\left[\|\mathbb{E}[c \mid w]\|_1\right] \leq \mathbb{E}\left[\mathbb{E}[\|c\|_1 \mid w]\right] = \mathbb{E}[\|c\|_1] < \infty$ by Assumption 1.

We will take $\delta := \delta_\ell(\cdot; \mathbb{P})$ defined in (EC.1), which is measurable by Lemma EC.5. For any $g : W \to \mathbb{R}^m$, and $w \in W$ such that $\mathbb{E}[L(g(w),c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] \geq \epsilon$, by $\mathbb{P}$-calibration and definition of δ_ℓ we have $\mathbb{E}[\ell(g(w),c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] \geq \delta_\ell(\epsilon, w; \mathbb{P})$, therefore the required property for δ is satisfied.

Applying the result of Steinwart (2007, Theorem 2.8) then gives the risk bound. □

EC.4. Proofs of Results from Section 4

Proof of Theorem 2. Denote

$$\begin{aligned} D_\ell(\alpha; w) &:= \left\{ d \in \mathbb{R}^m : \mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d,c) \mid w] \ < \ \alpha \right\} \\ D(\alpha; w) &:= \left\{ d \in \mathbb{R}^m : \mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d,c) \mid w] \ < \ \alpha \right\}. \end{aligned}$$

Note that

$$\operatorname*{arg\,min}_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] = \bigcap_{\alpha > 0} D_\ell(\alpha; w), \quad \operatorname*{arg\,min}_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] = \bigcap_{\alpha > 0} D(\alpha; w).$$

Suppose first that ℓ is $\mathbb{P}$-calibrated. Then for any $\epsilon > 0$, there exists $\delta > 0$ (which can depend on w) such that $D_\ell(\delta; w) \subseteq D(\epsilon; w)$. In particular, since $D_\ell(\alpha; w) \subseteq D_\ell(\alpha'; w)$ for $\alpha \leq \alpha'$, we have

$$\operatorname*{arg\,min}_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] = \bigcap_{0 < \alpha \leq \delta} D_\ell(\alpha; w) \subseteq D(\epsilon; w).$$

Taking the intersection of the right hand side over $\epsilon > 0$, we have

$$\operatorname*{arg\,min}_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] \subseteq \operatorname*{arg\,min}_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w],$$

hence ℓ is $\mathbb{P}$-Fisher consistent.

Suppose now that ℓ is not $\mathbb{P}$-calibrated. We show that it is also not $\mathbb{P}$-Fisher consistent. Fix an arbitrary $w \in W$. Note that the function $h : \mathbb{R}^m \to \mathbb{R}$ defined by $h(d) = \mathbb{E}[\ell(d, c) \mid w]$ is convex by convexity of $\ell(d, c)$, and hence under Assumption 1, it is continuous (see e.g., Rockafellar (1970, Theorem 10.1)).

Since ℓ is not $\mathbb{P}$-calibrated, there exists $w \in W$ and $\epsilon > 0$ such that for all $\delta > 0$, there exists $d(\delta) \in \mathbb{R}^m$ such that $h(d(\delta)) - \min_{d' \in \mathbb{R}^m} h(d') < \delta$ but $\mathbb{E}[L(d(\delta), c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d, c) \mid w] \geq \epsilon$.

Now, let $d_k = d(1/k)$ for $k \in \mathbb{N}$. Note that $\{d_k\}_{k \in \mathbb{N}} \subset D_\ell(1; w)$ which is compact since by Assumption 1 $\operatorname{arg\,min}_{d' \in \mathbb{R}^m} h(d')$ is compact, so all level sets are bounded (see, e.g., Rockafellar (1970, Corollary 8.7.1)). Therefore, there exists a convergent subsequence $d'_k \to d \in \operatorname{cl} D_\ell(1; w)$. Since h is continuous, we must have $d \in \operatorname{arg\,min}_{d' \in \mathbb{R}^m} h(d')$.

We now want to show that $d \notin \operatorname{arg\,min}_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w]$. We know from Lemma EC.1 that the argmin mapping $X^*(\cdot)$ is upper semi-continuous at d. Suppose for contradiction that $d \in \operatorname{arg\,min}_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w]$. Then we must have $X^*(d) \subseteq X^*(\mathbb{E}[c \mid w])$. Thus, for $\epsilon > 0$ the set

$$X^\circ(\epsilon') = \left\{ x' : f(x') + \mathbb{E}[c \mid w]^\top x' < \min_{x \in X} \left\{ f(x) + \mathbb{E}[c \mid w]^\top x \right\} + \epsilon' \right\}$$

is an "open" set (as $x \mapsto f(x) + \mathbb{E}[c \mid w]^\top x$ is continuous) containing $X^*(d)$. Note that this is not open in $\mathbb{R}^m$ by the usual topology, since $f(x)$ may be infinite for $x \notin X$. However, it is open when we work with $X \subset \mathbb{R}^m$ as the *entire* topological space with the induced topology from $\mathbb{R}^m$. Then, by Definition EC.1 of upper semi-continuity, there exists a neighbourhood $D^\circ(\epsilon')$ of d such that for any $d^\circ \in D^\circ(\epsilon')$, $X^*(d^\circ) \subset X^\circ(\epsilon')$, which means that $\mathbb{E}[L(d^\circ, c) \mid w] < \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] + \epsilon'$ since $x^*(d^\circ) \in X^*(d^\circ) \subseteq X^\circ(\epsilon')$.

But now consider $\epsilon' < \epsilon$. Since $d'_k \to d$, $D^\circ(\epsilon')$ is open, and $d \in D^\circ(\epsilon')$, we eventually have $d'_k \in D^\circ(\epsilon')$ for sufficiently large k. But this contradicts the fact that by construction of the sequence $\{d_k\}_{k \in \mathbb{N}}$ we have $\epsilon' < \epsilon < \mathbb{E}[L(d'_k, c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] = \mathbb{E}[c \mid w]^\top x^*(d'_k) - \min_{x \in X} \mathbb{E}[c \mid w]^\top x$. $\square$

Proof of Corollary 2. Fix some $\epsilon > 0$. Take $\delta > 0$ corresponding to ϵ in Corollary 1. Since $R_\ell(g_n, \mathbb{P}) \to R_\ell(\mathbb{P})$, we have $R_\ell(g_n, \mathbb{P}) \leq R_\ell(\mathbb{P}) + \delta$ eventually. By Theorem 1, we will also have $R(g_n, \mathbb{P}) \to R(\mathbb{P}) + \epsilon$ eventually. □

Proof of Example 10. Let us explore what $\arg\min_{d' \in \mathbb{R}} \mathbb{E}[\ell(d, c) \mid w]$ is for our setting. For convenience, we fix $w \in W$, and omit the w in the notation, so that $D^* = D^*_w$, $\mathbb{E}[\cdot] = \mathbb{E}[\cdot \mid w]$ and $\mathbb{P}[\cdot] = \mathbb{P}[\cdot \mid w]$. Then

$$2\mathbb{E}[\ell(d,c)] = \mathbb{E}[|2d - c|] - 2d\mathbb{E}[\text{sign}(c)] + \mathbb{E}[|c|] = \mathbb{E}[|2d-c|] + 2d\left(\mathbb{P}[c<0] - \mathbb{P}[c>0]\right) + \mathbb{E}[|c|].$$

This is a convex function in d, so we look at the subdifferential to determine its minimizers. Note that

$$\partial_d \mathbb{E}[|2d-c|] = \left\{2\left(\mathbb{P}[c<2d] - \mathbb{P}[c>2d] + s\mathbb{P}[c=2d]\right) : s \in [-1,1]\right\},$$

so

$$\partial_d \mathbb{E}[\ell(d,c)] = \left\{\mathbb{P}[c<2d] - \mathbb{P}[c>2d] + \mathbb{P}[c<0] - \mathbb{P}[c>0] + s\mathbb{P}[c=2d] : s \in [-1,1]\right\}.$$

For simplicity, let us assume that $\mathbb{P}[c = 2d] = 0$ for any d (many such distributions exist). Then $\mathbb{E}[\ell(d,c)]$ is differentiable with

$$\nabla_d \mathbb{E}[\ell(d,c)] = \mathbb{P}[c<2d] - \mathbb{P}[c>2d] + \mathbb{P}[c<0] - \mathbb{P}[c>0].$$

Denote d^* to be a minimizer of $\mathbb{E}[\ell(d,c)]$. If $\mathbb{P}[c<0] = \mathbb{P}[c>0]$, then setting $d=0$ gives $\nabla_d \mathbb{E}[\ell(d,c)] = 0$, so $d^* = c = 0$. If $\mathbb{P}[c<0] - \mathbb{P}[c>0] < 0$, then $\nabla_d \mathbb{E}[\ell(d,c)]|_{d=0} < 0$, so increasing d from 0 will decrease $\mathbb{E}[\ell(d,c)]$. Thus, $d^* > 0$. However, note that $\mathbb{P}[c<0] - \mathbb{P}[c>0] < 0$ implies that the median of c is also > 0. If $\mathbb{P}[c<0] - \mathbb{P}[c>0] > 0$, then $\nabla_d \mathbb{E}[\ell(d,c)]|_{d=0} > 0$, so decreasing d from 0 will decrease $\mathbb{E}[\ell(d,c)]$. Thus, $d^* < 0$. However, note that $\mathbb{P}[c<0] - \mathbb{P}[c>0] > 0$ implies that the median of c is also < 0. In all cases, the minimizer d^* is of the same sign as the median of c. Now, if $\mathbb{P}$ is a symmetric distribution, then the mean $\mathbb{E}[c]$ is equal to the median, and thus d^* has the same sign as $\mathbb{E}[c]$, so also minimizes $\mathbb{E}[L(d,c)]$. However, if the median has a different sign to the mean, then ℓ is not $\mathbb{P}$-Fisher consistent. Such distributions can be constructed by shifting a log-normal distribution, for example. □

Proof of Example 11. With the distribution $\mathbb{P}$ specified, $\min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell_{\text{SPO+}}(d', c)]$ can be expressed as the following linear program (making the change of variables $2d' \to d$):

$$\begin{aligned}
\min_{d,\gamma} \quad & \sum_{j \in [m]} p_j (d_j - \gamma_j) \\
\text{s.t.} \quad & \gamma_j \leq d_j, \ j \in [m] \\
& \gamma_j \leq d_k - 1, \ j,k \in [m], k \neq j \\
& d, \gamma \in \mathbb{R}^m.
\end{aligned}$$

We analyse this linear program. Fix a vector $d \in \mathbb{R}^m$. Let $j^* \in \arg\min_{j' \in [m]} d_{j'}$. Then since $p_k > 0$ for all $k \neq j^*$, the optimal choice of γ_k makes it as large as possible, so we set $\gamma_k = d_{j^*} - 1$ for $k \neq j^*$. In other words, for all but one index $j^* \in \arg\min_{j' \in [m]} d_{j'}$, we set $\gamma_j = \min_{j' \in [m]} d_{j'} - 1$. For j^*, we set $\gamma_{j^*} = \min\{d_{j^*}, \min_{j' \neq j^*} d_{j'} - 1\}$.

If there exists $j \neq j^*$ such that $d_{j^*} \leq d_j - 1$, then decreasing $d_j \downarrow d_{j^*} + 1$ does not violate any constraints since $\gamma_j = d_{j^*} - 1 < d_j$ and $\gamma_{j^*} = d_{j^*} \leq d_j - 1$, and decreases the objective. Therefore, without loss of generality, we assume that $d_j - 1 \leq d_{j^*}$ for all $j \neq j^*$. This implies that $\gamma_{j^*} = \min_{j' \neq j^*} d_{j'} - 1$.

Furthermore, if we have $j, k \in [m] \setminus \{j^*\}$, $j \neq k$ such that $d_j < d_k$, note that we can decrease $d_k \downarrow d_j$ without violating any constraints, since $\gamma_{j'} = d_{j^*} - 1 \leq d_j - 1 < d_k - 1 < d_k$ for all $j' \neq j^*$ and $\gamma_{j^*} \leq d_j - 1 < d_k - 1$. This implies that, without loss of generality, we can assume that for $j \neq j^*$, we have $d_j = \delta$ for some $\delta \in [d_{j^*}, d_{j^*} + 1]$. In particular, this implies that $\gamma_{j^*} = \delta - 1$, thus the objective becomes

$$\sum_{j \in [m]} p_j (d_j - \gamma_j) = (\delta - d_{j^*} + 1) \sum_{j \neq j^*} p_j + p_{j^*}(d_{j^*} - \delta + 1) = (1 - 2p_{j^*})(\delta - d_{j^*}) + 1.$$

This shows that if $p_{j^*} > 1/2$, then we should make δ as large as possible, i.e., $\delta = d_{j^*} + 1$. On the other hand, when $p_{j^*} < 1/2$, we set $\delta = d_{j^*}$, i.e., the optimal vector d^* is constant.

This implies that, if there exists $j^* \in [m]$ such that $p_{j^*} > 1/2$, and necessarily $j^* = \arg\max_{j' \in [m]} p_{j'}$, then the minimizers of $\mathbb{E}[\ell(d, c)]$ take the form $d_\alpha = (\alpha \mathbf{1}_m - e_{j^*})/2$ for $\alpha \in \mathbb{R}$. Clearly, $\arg\min_{j' \in [m]} d_{\alpha, j'} = j^*$, so for such distributions $\mathbb{P}$, $\ell_{\text{SPO+}}$ is $\mathbb{P}$-Fisher consistent.

On the other hand, for distributions $\mathbb{P}$ with $\max_{j' \in [m]} p_{j'} < 1/2$, $\ell_{\text{SPO+}}$ is not $\mathbb{P}$-Fisher consistent, since the set of minimizers of $\mathbb{E}[\ell(d, c)]$ are the vectors $d_\alpha = \alpha \mathbf{1}_m$, $\alpha \in \mathbb{R}$, which cannot in general pick out the maximum probability class $j \in [m]$, i.e., the highest p_j. □

EC.5. Proofs of Results from Section 5

Proof of Lemma 2. The "only if" direction was established in Remark 6, so we only need to prove the "if" direction.

When $\delta_\ell(\epsilon; \mathcal{P}) > 0$, take $0 < \delta \leq \delta_\ell(\epsilon; \mathcal{P})$, and noting that $\delta_\ell(\cdot; \mathcal{P})$ is non-decreasing, we get for any $d \in \mathbb{R}^m$, $w \in W$ and $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}[\ell(d, c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] \leq \delta < \delta_\ell(\epsilon; \mathcal{P}).$$

If $d \in \mathbb{R}^m$, $w \in W$ and $\mathbb{P} \in \mathcal{P}$ were such that $\mathbb{E}[L(d, c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] > \epsilon$, we reach a contradiction since we would then by definition of $\delta_\ell(\cdot; \mathcal{P})$ in (14) have $\mathbb{E}[\ell(d, c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] \geq \delta_\ell(\epsilon; \mathcal{P})$. Thus, for any $w \in W$ and $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}[\ell(d, c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w] \leq \delta \implies \mathbb{E}[L(d, c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w] \leq \epsilon.$$

□

Proof of Theorem 3. When ℓ is $\mathcal{P}$-uniformly calibrated, we know that $\delta_\ell(\epsilon;\mathcal{P})>0$ for any $\epsilon>0$. Steinwart (2007, Lemma A.6) shows that this implies $\delta^{**}(\epsilon;\mathcal{P})>0$ for $\epsilon\in(0,B_f+B_CB_X]$.

We can now utilize Steinwart (2007, Theorem 2.13) to derive the risk bound. In order to do so, note that Steinwart (2007, Theorem 2.13) requires us to verify that for any $g:W\to\mathbb{R}$,

$$\operatorname*{ess\,sup}_{w\in W}\mathbb{E}[L(g(w),c)\mid w]-\min_{d'\in\mathbb{R}^m}\mathbb{E}[L(d',c)\mid w]\le B_f+B_CB_X,$$

where ess sup stands for essential supremum. The relation above follows from Remark 5 and from the definition of $\delta_\ell(\cdot;\mathcal{P})$ that ensures that for any $\epsilon>0$ and $w\in W$, we have

$$\left\{g:\mathbb{E}[\ell(g(w),c)\mid w]-\min_{d'\in\mathbb{R}^m}\mathbb{E}[\ell(d',c)\mid w]<\delta_\ell(\epsilon;\mathcal{P})\right\}\subseteq\left\{g:\mathbb{E}[L(g(w),c)\mid w]-\min_{d'\in\mathbb{R}^m}\mathbb{E}[L(d',c)\mid w]<\epsilon\right\}.$$

□

Proof of Lemma 3. Fix an arbitrary $w\in W$. Note that $\mathbb{E}[c\mid w]\in\{c':x^*(c')\in X^*(\mathbb{E}[c\mid w])\}=\arg\min_{d'\in\mathbb{R}^m}\mathbb{E}[L(d',c)\mid w]$, hence we have

$$\begin{aligned}\mathbb{E}[L(d,c)\mid w]-\min_{d'\in\mathbb{R}^m}\mathbb{E}[L(d',c)\mid w]&=f(x^*(d))+\mathbb{E}[c\mid w]^\top x^*(d)-\min_{d'\in\mathbb{R}^m}\left\{f(x^*(d'))+\mathbb{E}[c\mid w]^\top x^*(d')\right\}\\&=f(x^*(d))-f(x^*(\mathbb{E}[c\mid w]))+\mathbb{E}[c\mid w]^\top(x^*(d)-x^*(\mathbb{E}[c\mid w])).\end{aligned}$$

Hence

$$\begin{aligned}\delta_\ell(\epsilon;\mathcal{P})&=\inf_{\substack{d\in\mathbb{R}^m\\w\in W\\\mathbb{P}\in\mathcal{P}}}\left\{\mathbb{E}[\ell(d,c)\mid w]-\min_{d'\in\mathbb{R}^m}\mathbb{E}[\ell(d',c)\mid w]:\mathbb{E}[L(d,c)\mid w]-\min_{d'\in\mathbb{R}^m}\mathbb{E}[L(d',c)\mid w]\ \ge\ \epsilon\right\}\\&=\inf_{d,\bar c\in\mathbb{R}^m}\ \inf_{\substack{w\in W\\\mathbb{P}\in\mathcal{P}\\\mathbb{E}[c|w]=\bar c}}\left\{\mathbb{E}[\ell(d,c)\mid w]-\min_{d'\in\mathbb{R}^m}\mathbb{E}[\ell(d',c)\mid w]:f(x^*(d))-f(x^*(\bar c))+\bar c^\top(x^*(d)-x^*(\bar c))\ \ge\ \epsilon\right\}\\&=\inf_{x,x'\in X}\ \inf_{\substack{d:x^*(d)=x\\\bar c:x^*(\bar c)=x'}}\ \inf_{\mathbb{P}:\mathbb{E}[c]=\bar c}\left\{\mathbb{E}[\ell(d,c)]-\min_{d'\in\mathbb{R}^m}\mathbb{E}[\ell(d',c)]:f(x)-f(x')+\bar c^\top(x-x')\ \ge\ \epsilon\right\}.\end{aligned}$$

□

Proof of Lemma 4. Fix arbitrary distinct $x,x'\in X$. Consider the halfspace

$$H_0(x,x')=\left\{d':f(x)-f(x')+(d')^\top(x-x')\le0\right\}\supseteq\{d':x^*(d')=x\}.$$

Since $x^*(d)=x$, we have $d\in H_0(x,x')$. Now, if $\bar c\in H_0(x,x')$, then $f(x)-f(x')+\bar c^\top(x-x')\le0$, hence $\|d-\bar c\|_2\ge0$.

On the other hand, if $x^*(\bar c)=x'$ and $f(x)-f(x')+\bar c^\top(x-x')>0$, then we have $\bar c\notin H_0(x,x')$, hence the distance between $\bar c$ and d is bounded below by the distance between $\bar c$ and the halfspace $H_0(x,x')$, which has the expression

$$\|d-\bar c\|_2\ge\inf_{d'\in H_0(x,x')}\|d'-\bar c\|_2=\frac{f(x)-f(x')+\bar c^\top(x-x')}{\|x-x'\|_2}.$$

□

Proof of Lemma 5. The usual bias-variance decomposition for squared error gives us

$$\begin{aligned}\mathbb{E}[\ell_{\mathrm{LS}}(d,c)\mid w] &= \mathbb{E}[\|d-c\|_2^2 \mid w]\\ &= \|d-\mathbb{E}[c\mid w]\|_2^2 + 2\mathbb{E}\left[(d-\mathbb{E}[c\mid w])^\top(\mathbb{E}[c\mid w]-c)\right] + \mathbb{E}\left[\|\mathbb{E}[c\mid w]-c\|_2^2\right]\\ &= \|d-\mathbb{E}[c\mid w]\|_2^2 + \mathbb{E}\left[\|\mathbb{E}[c\mid w]-c\|_2^2\right].\end{aligned}$$

Hence, we can minimize this by choosing $d=\mathbb{E}[c\mid w]$, and

$$\min_{d'\in\mathbb{R}^m}\mathbb{E}[\ell_{\mathrm{LS}}(d',c)\mid w] = \mathbb{E}\left[\|\mathbb{E}[c\mid w]-c\|_2^2\right].$$

Therefore,

$$\mathbb{E}[\ell_{\mathrm{LS}}(d,c)\mid w] - \min_{d'\in\mathbb{R}^m}\mathbb{E}[\ell_{\mathrm{LS}}(d',c)\mid w] = \|d-\mathbb{E}[c\mid w]\|_2^2.$$

Substituting this into (15) and using the fact that the definition of $\mathcal{P}$ tells us that $\bar{c}=\mathbb{E}[c\mid w]$ can take on any point in $\mathrm{Conv}(C)$ gives the result. □

Proof of Theorem 4. Fixing distinct $x,x'\in X$, notice that if $\bar{c},d$ are chosen according to the conditions of Lemma 4, together with the condition that $f(x)-f(x')+\bar{c}^\top(x-x')>\epsilon$, then $\|d-\bar{c}\|_2>\epsilon/\|x-x'\|_2\geq\epsilon/B_X>0$. Together with Lemma 5, we have, for all $\epsilon>0$,

$$\delta_{\ell_{\mathrm{LS}}}(\epsilon;\mathcal{P})\geq\frac{\epsilon^2}{B_X^2}>0.$$

Then $\mathcal{P}$-uniform calibration follows from Lemma 2. □

Proof of Corollary 3. The result follows by observing that $\epsilon^2/B_X^2\leq\delta^{**}(\epsilon)$ since $\epsilon\mapsto\epsilon^2/B_X^2$ is already convex, and then applying Theorem 3. □

Proof of Theorem 5. Analogous to (14), define

$$\delta_j(\epsilon;\mathcal{P}_{\mathrm{sym}}) := \inf_{\substack{d_j\in\mathbb{R}\\ w\in W\\ \mathbb{P}\in\mathcal{P}_{\mathrm{sym}}}}\left\{\mathbb{E}[\ell_j(d_j,c_j)\mid w]-\min_{d_j'\in\mathbb{R}}\mathbb{E}[\ell_j(d_j',c_j)\mid w] : (d_j-\mathbb{E}[c_j\mid w])^2>\epsilon\right\}.$$

We first show that $\delta_j(\epsilon;\mathcal{P}_{\mathrm{sym}})>0$ for all $\epsilon>0$.

First, fix $\mathbb{P}\in\mathcal{P}_{\mathrm{sym}}$ and $w\in W$, and observe that for any d_j

$$\begin{aligned}\mathbb{E}[\ell_j(\mathbb{E}[c_j\mid w]+d_j,c_j)\mid w] &= \mathbb{E}[\psi_j(d_j-(c_j-\mathbb{E}[c_j\mid w])\mid w] = \mathbb{E}[\psi_j(d_j-(\mathbb{E}[c_j\mid w]-c_j))\mid w]\\ &= \mathbb{E}[\psi_j(-d_j+\mathbb{E}[c_j\mid w]-c_j)\mid w]\\ &= \mathbb{E}[\ell_j(\mathbb{E}[c_j\mid w]-d_j,c_j)\mid w].\end{aligned}$$

Since ψ is strictly convex, $\mathbb{E}[\ell_j(d_j,c_j)\mid w]$ is strictly convex in d_j, thus for any $d_j\neq 0$,

$$\begin{aligned}&\mathbb{E}[\ell_j(\mathbb{E}[c_j\mid w]+d_j,c_j)\mid w]-\mathbb{E}[\ell_j(\mathbb{E}[c_j\mid w],c_j)\mid w]\\ &=\frac{1}{2}\mathbb{E}[\ell_j(\mathbb{E}[c_j\mid w]+d_j,c_j)\mid w]+\frac{1}{2}\mathbb{E}[\ell_j(\mathbb{E}[c_j\mid w]-d_j,c_j)\mid w]-\mathbb{E}[\ell_j(\mathbb{E}[c_j\mid w],c_j)\mid w]\\ &\geq\frac{1}{4}\mathbb{E}\left[\delta_j(4d_j^2)\mid w\right]=\frac{1}{4}\delta_j(4d_j^2).\end{aligned}$$

This shows that $\delta_j(\epsilon;\mathcal{P}_{\text{sym}}) \geq \delta_j(2\epsilon)/4$. Now, following the outline in Section 5.1 and proceeding similarly to the proof of Theorem 3, we deduce the risk bound. □

Proof of Proposition 1. The proof is by construction. We will fix the mean of our class to be ϵ. Let ϕ be the density function of the standard normal distribution, and Φ be the distribution function (note that $\phi(c-\epsilon)$ is the density function of a $N(\epsilon,1)$ random variable. Let $z_\epsilon = \Phi(-\epsilon) + 1 - \Phi(\epsilon)$ denote the probability that a standard normal variable is $< -\epsilon$ or $> \epsilon$. Furthermore, let $\{h(\cdot;\alpha)\}_{\alpha\in(0,1)}$ be a class of continuous functions such that for each $\alpha \in (0,2/3)$, $h(r;\alpha) > 0$ for $r \in [0,1]$, $h(1;\alpha) = 1$, $\int_{r=0}^{1} h(r;\alpha)dr = \alpha$. Such a class can be defined as follows:

$$h(r;\alpha) = \begin{cases} \alpha/2, & 0 \leq r \leq (2-3\alpha)/(2-\alpha) \\ (2-\alpha)^2(r-1)/(4\alpha)+1, & (2-3\alpha)/(2-\alpha) < r \leq 1. \end{cases}$$

For each $k \in \mathbb{N}$, define the following density function $\psi^{(k)}$:

$$\psi^{(k)}(c) = \begin{cases} \dfrac{z_\epsilon + (1-1/k)(1-z_\epsilon)}{z_\epsilon}\phi(c-\epsilon), & c \leq 0 \\ \dfrac{z_\epsilon + (1-1/k)(1-z_\epsilon)}{z_\epsilon}\phi(-\epsilon)h\left(1-c/\epsilon; \dfrac{(1-z_\epsilon)z_\epsilon}{2k\epsilon\left(z_\epsilon+(1-1/k)(1-z_\epsilon)\right)\phi(-\epsilon)}\right), & 0 < c < \epsilon \\ \dfrac{z_\epsilon + (1-1/k)(1-z_\epsilon)}{z_\epsilon}\phi(-\epsilon)h\left(c/\epsilon-1; \dfrac{(1-z_\epsilon)z_\epsilon}{2k\epsilon\left(z_\epsilon+(1-1/k)(1-z_\epsilon)\right)\phi(-\epsilon)}\right), & \epsilon \leq c < 2\epsilon \\ \dfrac{z_\epsilon + (1-1/k)(1-z_\epsilon)}{z_\epsilon}\phi(\epsilon-c), & c \geq 2\epsilon. \end{cases}$$

By construction, $\psi^{(k)}(c)$ is continuous and positive for all $c \in \mathbb{R}$, and integrates to 1. Let $\mathbb{P}^{(k)}$ denote the corresponding probability distribution, and by construction we have $\mathbb{P}^{(k)}[0 \leq c \leq 2\epsilon] = (1-z_\epsilon)/k$. Therefore $\mathbb{P}^{(k)}[c>0] - \mathbb{P}^{(k)}[c<0] = (1-z_\epsilon)/k \to 0$ as $k \to \infty$, but $\mathbb{E}^{(k)}[c] = \epsilon$ since $\psi^{(k)}(c-\epsilon) = \psi(\epsilon - c)$ is symmetric about ϵ. □

Proof of Proposition 2. We know that

$$\delta_{\ell_{\text{SPO+}}}(\epsilon;\mathcal{P}_{\text{cont,sym},\alpha}) = \inf_{w\in W} \inf_{\substack{\mathbb{P}\in\mathcal{P}_{\text{cont,sym},\alpha} \\ |\mathbb{E}[c|w]|>\epsilon}} \left\{\mathbb{E}[c \mid w]\left(\mathbb{P}[c>0 \mid w] - \mathbb{P}[c<0 \mid w]\right)\right\}.$$

Using the property of $\mathcal{P}_{\text{cont,sym},\alpha}$, we deduce that

$$\delta_{\ell_{\text{SPO+}}}(\epsilon;\mathcal{P}_{\text{cont,sym},\alpha}) \geq \inf_{w\in W} \inf_{\substack{\mathbb{P}\in\mathcal{P}_{\text{cont,sym},\alpha} \\ |\mathbb{E}[c|w]|>\epsilon}} \alpha|\mathbb{E}[c \mid w]| \geq \alpha\epsilon.$$

Thus since $\delta_{\ell_{\text{SPO+}}}(\epsilon;\mathcal{P}_{\text{cont,sym},\alpha}) > 0$ for any $\epsilon > 0$, we have uniform calibration by Lemma 2. Furthermore, Theorem 3 gives us the risk bound. □

EC.6. Proof of Results from Section 6

Proof of Proposition 3. Using Lagrange duality we know that the dual problem is

$$\min_{\gamma}\left\{b\gamma + \frac{1}{2}(d-\gamma p)^\top Q^{-1}(d-\gamma p)\right\} = -\min_{x}\left\{\frac{1}{2}x^\top Qx - d^\top x : p^\top x = b\right\},$$

and the optimal solution is $x^*(d) = Q^{-1}(d - \gamma^* p)$ where γ^* is the optimal dual solution. The closed form solution is $\gamma^* = \frac{1}{p^\top Q^{-1} p}\left(p^\top Q^{-1} d - b\right)$, hence

$$x^*(d) = Q^{-1}d - \frac{1}{p^\top Q^{-1}p}\left(p^\top Q^{-1}d - b\right)Q^{-1}p = Ad + \frac{b}{p^\top Q^{-1}p}Q^{-1}p.$$

Observe that $Ap = 0$, so

$$\begin{aligned}
x^*(d)^\top Q x^*(d) &= \left(Ad + \frac{b}{p^\top Q^{-1}p}Q^{-1}p\right)^\top \left(QAd + \frac{b}{p^\top Q^{-1}p}p\right) \\
&= d^\top A^\top QAd + \frac{b^2}{p^\top Q^{-1}p} \\
&= d^\top A^\top \left(I - \frac{p(Q^{-1}p)^\top}{p^\top Q^{-1}p}\right)d + \frac{b^2}{p^\top Q^{-1}p} \\
&= d^\top Ad + \frac{b^2}{p^\top Q^{-1}p} \\
c^\top x^*(d) &= c^\top Ad + \frac{b \cdot p^\top Q^{-1}c}{p^\top Q^{-1}p} \\
\frac{1}{2}x^*(d)^\top Q x^*(d) - c^\top x^*(d) &= \frac{1}{2}d^\top Ad - c^\top Ad + \frac{b^2/2 - b\cdot p^\top Q^{-1}c}{p^\top Q^{-1}p}.
\end{aligned}$$

Clearly we have $\frac{1}{2}x^*(c)^\top Q x^*(c) - c^\top x^*(d) = -\frac{1}{2}c^\top Ac + \frac{b^2/2 - b\cdot p^\top Q^{-1}c}{p^\top Q^{-1}p}$ so therefore

$$L(d,c) = \frac{1}{2}d^\top Ad - c^\top Ad + \frac{1}{2}c^\top Ac = \frac{1}{2}(d-c)^\top A(d-c).$$

The result now follows. □

Proof of Proposition 4. First, notice that $\mathbb{E}\left[\frac{1}{2}\|Vw - c\|_2^2\right] = \frac{1}{2}\operatorname{Tr}(V^\top V\mathbb{E}[ww^\top]) - \operatorname{Tr}(V^\top \mathbb{E}[cw^\top]) + \frac{1}{2}\mathbb{E}[\|c\|_2^2]$. Via standard vector calculus, we have $\nabla_V \mathbb{E}\left[\frac{1}{2}\|Vw - c\|_2^2\right] = V\mathbb{E}[ww^\top] - \mathbb{E}[cw^\top]$, therefore the optimality condition of the least squares predictor is

$$V\mathbb{E}[ww^\top] = \mathbb{E}[cw^\top].$$

Now observe that we can write $\mathbb{E}\left[\frac{1}{2}(Vw-c)^\top A(Vw-c)\right] = \frac{1}{2}\operatorname{Tr}(V^\top AV\mathbb{E}[ww^\top]) - \operatorname{Tr}(V^\top\mathbb{E}[Acw^\top]) + \frac{1}{2}\mathbb{E}[c^\top Ac]$. The gradient is $\nabla_V \mathbb{E}\left[\frac{1}{2}(Vw-c)^\top A(Vw-c)\right] = AV\mathbb{E}[ww^\top] - A\mathbb{E}[cw^\top]$. The optimality condition is $AV\mathbb{E}[ww^\top] = A\mathbb{E}[cw^\top]$. We can alternatively represent this as

$$V\mathbb{E}[ww^\top] \in \mathbb{E}[cw^\top] + \left\{\tilde{V} : A\tilde{V} = \mathbf{0}\right\}.$$

Since $\mathbb{E}[ww^\top]$ is invertible, the result follows. □

Proof of Proposition 5 This result follows immediately from Lemma EC.6 below. □

LEMMA EC.6. *Assume that* $\|d\|_1 \le M$ *and that* $M_\tau = M/\left(\min_{j\in[n]} p_j\right)$. *The set of optimal solutions to* $\max_{x\in X}\left\{d^\top x - \frac{\lambda}{2}\|x\|_2^2\right\}$ *can be characterized as*

$$\left\{ x : \begin{array}{c} p^\top x \le B,\ \mathbf{0}\le x\le \mathbf{1} \\ \tau\ge 0,\ q,z\in\{0,1\}^n,\ v\in\{0,1\} \\ \tau \le M_\tau v,\ B - p^\top x \le B(1-v) \\ d_j - p_j\tau \le Mq_j,\ p_j\tau - d_j \le (M_\tau p_j + M)(1-q_i),\ j\in[m] \\ d_j - p_j\tau - \lambda \le Mz_i,\ \lambda + p_j\tau - d_j \le (M_\tau p_j + M + \lambda)(1-z_j),\ j\in[m] \\ x_j \le q_j,\ x_j \ge z_j,\ j\in[m] \\ \lambda x_j \le d_j - p_j\tau + (M + M_\tau p_j)(1-q_j),\ \lambda x_j \ge d_j - p_j\tau - Mz_j,\ j\in[m]. \end{array} \right\}.$$

Proof of Lemma EC.6. Fix d. We consider the primal-dual pair of problems for the regularized fractional knapsack:

$$\begin{aligned} &\max_x \left\{ d^\top x - \frac{\lambda}{2}\|x\|_2^2 : p^\top x \le B,\ \mathbf{0}\le x\le \mathbf{1} \right\} \\ &= \min_{s,y,\tau} \left\{ B\tau + \mathbf{1}^\top y + \frac{1}{2\lambda}\|s\|_2^2 : s \ge d - p\tau - y,\ s,y,\tau \ge 0 \right\}. \end{aligned}$$

Using the complementary slackness conditions, the set of primal-dual optimal pairs (x,s,y,τ) can be written as

$$X^*(d) = \left\{ (x,s,y,\tau) : \begin{array}{c} p^\top x \le B,\ \mathbf{0}\le x\le \mathbf{1}, \\ s \ge d - p\tau - y,\ s,y,\tau \ge 0, \\ \tau(B - p^\top x) = 0, \\ y_i(1-x_i) = 0,\ i\in[n] \\ x_i(s_i - (d_i - p_i\tau - y_i)) = 0,\ i\in[n] \\ s = \lambda x \end{array} \right\}$$

$$\mathrm{Proj}_{x,\tau}(X^*(d)) = \left\{ (x,\tau) : \begin{array}{c} p^\top x \le B,\ \mathbf{0}\le x\le \mathbf{1}, \\ \tau \ge 0, \\ \tau(B - p^\top x) = 0, \\ \lambda x_i = \max\{0, \min\{\lambda, d_i - p_i\tau\}\},\ i\in[n] \end{array} \right\}.$$

To see why the second equality holds, consider some solution $(x,s,y,\tau)\in X^*(d)$. If $d_i - p_i\tau - y_i > 0$, then we need $\lambda x_i = s_i = d_i - p_i\tau - y_i$, which follows from $s_i \ge d_i - p_i\tau - y_i$, $x_i(s_i - (d_i - p_i\tau - y_i)) = 0$ and $s_i = \lambda x_i$. If $d_i - p_i\tau - y_i \le 0$, then since $s_i = \lambda x_i$, we would have $x_i(s_i - (d_i - p_i\tau - y_i)) > 0$ if $s_i = \lambda x_i > 0$, so we must have $s_i = \lambda x_i = 0$. Therefore $\lambda x_i = \max\{0, d_i - p_i\tau - y_i\}$. We now show that $y_i = \max\{0, d_i - p_i\tau - \lambda\}$. To see this, suppose that $d_i - p_i\tau - \lambda > 0$. We know that $y_i \ge d_i - p_i\tau -$

$\lambda x_i > 0$, and since $y_i(1-x_i)=0$, we have $x_i=1$. Since $\lambda x_i = \lambda = \max\{0, d_i - p_i\tau - y_i\} = d_i - p_i\tau - y_i$ implies that $y_i = d_i - p_i\tau - \lambda$. Now suppose that $d_i - p_i\tau - \lambda \le 0$. If $y_i > 0$, then since $y_i(1-x_i)=0$, we necessarily have $x_i = 1$. But then $\lambda x_i = \lambda \le d_i - p_i\tau - y_i$ is a contradiction. Therefore we necessarily have $y_i = 0$. Substituting $y_i = \max\{0, d_i - p_i\tau - \lambda\}$ into $\lambda x_i = \max\{0, d_i - p_i\tau - y_i\}$ gives us $\lambda x_i = \max\{0, \min\{\lambda, d_i - p_i\tau\}\}$.

We assume that $\|d\|_1 \le M$, and that we are given an a priori bound $\tau \le M_\tau$. We can model the constraint $\tau(B - p^\top x) = 0$ as

$$\tau \le M_\tau v, \quad B - p^\top x \le B(1-v), \quad v \in \{0,1\}.$$

We now describe how to model the constraint $\lambda x_i = \max\{0, \min\{\lambda, d_i - p_i\tau\}\}$. First, since $\lambda > 0$, we have that $\min\{\lambda, d_i - p_i\tau\} \ge 0$ if and only if $d_i - p_i\tau \ge 0$. Let q_i be an indicator variable for this event, which we model as

$$d_i - p_i\tau \le Mq_i, \quad p_i\tau - d_i \le (M_\tau p_i + M)(1-q_i).$$

Let z_i be an indicator variable for the event $d_i - p_i\tau \ge \lambda$, so we need the constraints

$$d_i - p_i\tau - \lambda \le Mz_i, \quad \lambda + p_i\tau - d_i \le (M_\tau p_i + M + \lambda)(1-z_i).$$

Note that implicitly, we have $q_i \ge z_i$. When $q_i = 0$, we have $\lambda x_i = 0$. When $q_i = 1$ and $z_i = 1$, we have $x_i = 1$, and when $q_i = 1$ and $z_i = 0$, we have $\lambda x_i = d_i - p_i\tau$. Therefore we need the constraints

$$x_i \le q_i, \quad x_i \ge z_i, \quad \lambda x_i \le d_i - p_i\tau + (M + M_\tau p_i)(1-q_i), \quad \lambda x_i \ge d_i - p_i\tau - Mz_i.$$

This shows that the proposed MIP representation is correct. □

Proof of Proposition 6. The dual of $\min_{x\in X}\left(2d_j x_j + \sum_{j'\in[m], j'\neq j}(2d_{j'} - 1)x_j\right)$ is $\max_\gamma \{\gamma : \gamma \le 2d_j,\ \gamma \le 2d_{j'} - 1,\ j' \in [m]\setminus\{j\}\}$. The lifted representation immediately follows from this. □