A Omitted Proofs

We define the multicalibration error of $\tilde{p}$ wrt $\mathcal{C}$ under $\mathcal{D}$ as

$$\mathsf{MCE}_{\mathcal{D}}(f, \mathcal{C}) = \max_{c \in \mathcal{C}} \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [c(\mathbf{x})(\mathbf{y} - \mathbf{v})] \right| \right].$$

We define the swap multicalibration error of $\tilde{p}$ wrt $\mathcal{C}$ under $\mathcal{D}$ as

$$\mathsf{sMCE}_{\mathcal{D}}(\tilde{p}, \mathcal{C}) = \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\max_{c \in \mathcal{C}} \left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [c(\mathbf{x})(\mathbf{y} - \mathbf{v})] \right| \right]$$

A.1 Properties of Swap Notions of Supervised Learning

Proof of Claim 2.4. We let $\ell_v = \ell$ for all $v \in \mathrm{Im}(\tilde{p})$, so that $k(v) = k_\ell(v)$. We pick the hypothesis

$$h_v = \arg\min_{h \in \mathcal{H}} \mathop{\mathbf{E}}_{\mathcal{D}|_v} [\ell(\mathbf{y}, h(\mathbf{x}))]$$

The swap omniprediction guarantee reduces to

$$\mathop{\mathbf{E}}_{\mathbf{v} \in \mathcal{D}_{\tilde{p}}} [\mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell(\mathbf{y}, k_\ell(\mathbf{v}))] = \mathop{\mathbf{E}}_{\mathcal{D}}[\ell(\mathbf{y}, k_\ell(\tilde{p}(\mathbf{x})))] \le \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \min_{h \in \mathcal{H}} \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell(\mathbf{y}, h(\mathbf{x}))] + \delta.$$

This implies that $f = k_\ell \circ \tilde{p}$ is a swap agnostic learner for every $\ell \in \mathcal{L}$ since we allow the choice of h to depend on $\tilde{p}(\mathbf{x})$ which is more informative than $f(\mathbf{x}) = k_\ell(\tilde{p}(\mathbf{x}))$. ■

Proof of Claim 2.7. We have

$$\begin{aligned} \mathsf{sMCE}_{\mathcal{D}}(\tilde{p}, \mathcal{C}) &= \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\max_{c \in \mathcal{C}} \left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [c(\mathbf{x})(\mathbf{y}^* - \mathbf{v})] \right| \right] \\ &\ge \max_{c \in \mathcal{C}} \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [c(\mathbf{x})(\mathbf{y} - \mathbf{v})] \right| \right] = \mathsf{MCE}_{\mathcal{D}}(\tilde{p}, \mathcal{C}) \end{aligned}$$

since the expectation of the max is higher than the max of expectations. Bounding the RHS by α is equivalent to $(\mathcal{C}, \alpha)$-multicalibration. ■

Proof of Claim 2.9. The ℓ_∞ bound is immediate from the definition of $\bar{p}$. We bound the swap multicalibration error of tf. We have $\bar{p}(\mathbf{x}) = j\delta$ iff $\tilde{p}(\mathbf{x}) \in B_j$, so that $|\tilde{p}(\mathbf{x}) - j\delta| \le \delta$ holds conditioned on this event. So

$$\begin{aligned} \mathsf{sMCE}_{\mathcal{D}}(\bar{p}, \mathcal{C}) &= \sum_{j \in [m]} \Pr[\bar{p}(\mathbf{x}) = j\delta] \max_{c \in \mathcal{C}} \left| \mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})(\mathbf{y} - j\delta) | \bar{p}(\mathbf{x}) = j\delta] \right| \\ &= \sum_{j \in [m]} \Pr[\tilde{p}(\mathbf{x}) \in B_j] \max_{c \in \mathcal{C}} \left| \mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})(\mathbf{y} - j\delta) | \tilde{p}(\mathbf{x}) \in B_j] \right| \\ &\le \sum_{j \in [m]} \Pr[\tilde{p}(\mathbf{x}) \in B_j] \left(\delta + \max_{c \in \mathcal{C}} \left| \mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})(\mathbf{y} - \tilde{p}(\mathbf{x})) | \tilde{p}(\mathbf{x}) \in B_j] \right| \right) \\ &\le \delta + \sum_{j \in [m]} \Pr[\tilde{p}(\mathbf{x}) \in B_j] \max_{c \in \mathcal{C}} \left| \mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})(\mathbf{y} - \tilde{p}(\mathbf{x})) | \tilde{p}(\mathbf{x}) \in B_j] \right| \end{aligned} \tag{15}$$

Let us fix a bucket B_j and a particular $c \in \mathcal{C}$. For $\beta \ge \alpha$ to be specified later we have

$$\begin{aligned} |\mathbf{E}[c(\mathbf{x})(\mathbf{y} - \tilde{p}(\mathbf{x})) | \tilde{p}(\mathbf{x}) \in B_j]| &\le \Pr[c(\mathbf{x})(\mathbf{y} - \tilde{p}(\mathbf{x})) \ge \beta | \tilde{p}(\mathbf{x}) \in B_j] + \beta \Pr[c(\mathbf{x})(\mathbf{y} - f(x)) \le \beta | \tilde{p}(\mathbf{x}) \in B_j] \\ &\le \frac{\Pr[\tilde{p}(\mathbf{x}) \in \mathrm{Bad}_\beta(c, f) \cap B_j]}{\Pr[\tilde{p}(\mathbf{x}) \in B_j]} + \beta \\ &\le \frac{\Pr[\tilde{p}(\mathbf{x}) \in \mathrm{Bad}_\beta(c, f)]}{\Pr[\tilde{p}(\mathbf{x}) \in B_j]} + \beta \\ &\le \frac{\alpha/\beta}{\Pr[\tilde{p}(\mathbf{x}) \in B_j]} + \beta. \end{aligned}$$

Since this bound holds for every c, it holds for the max over $c \in \mathcal{C}$ conditioned on $\tilde{p}(\mathbf{x}) \in B_j$. Hence

$$\begin{aligned}\sum_{j\in[m]} \Pr[\tilde{p}(\mathbf{x}) \in B_j] \max_{c\in\mathcal{C}} \left|\mathop{\mathbf{E}}_{\mathcal{D}}[c(\mathbf{x})(\mathbf{y}-\tilde{p}(\mathbf{x}))|\tilde{p}(\mathbf{x}) \in B_j]\right| &\le \sum_{j\in[m]} \Pr[\tilde{p}(\mathbf{x}) \in B_j] \left(\frac{\alpha/\beta}{\Pr[\tilde{p}(\mathbf{x}) \in B_j]} + \beta\right) \\ &\le \frac{\alpha}{\beta\delta} + \beta,\end{aligned}$$

where we use $m = 1/\delta$. Plugging this back into Equation (15) gives

$$\mathsf{sMCE}_{\mathcal{D}}(\bar{p}, \mathcal{C}) = \mathop{\mathbf{E}}_{\mathbf{v}\sim\bar{p}_{\mathcal{D}}}\left[\max_{c\in\mathcal{C}}\left|\mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}}[c(\mathbf{x})(\mathbf{y}-\mathbf{v})]\right|\right] \le \frac{\alpha}{\beta\delta} + \beta + \delta.$$

Taking $\beta = \sqrt{\alpha/\delta}$ gives the desired claim. ■

A.2 Omitted Proofs from Main Result

Proof of Lemma 3.2. We will show that for $p, p' \in [0,1]$ and $t_0 \in I_\ell$, we have

$$\ell(p, t_0) - \ell(p', t_0) \le |p - p'|B.$$

By the definition of $\ell(p,t)$, we have

$$\begin{aligned}\ell(p, t_0) - \ell(p', t_0) &= (p - p')\ell(0, t_0) + (1 - p - 1 + p')\ell(1, t_0) \\ &= (p - p')(\ell(0, t_0) - \ell(1, t_0))\end{aligned}$$

Taking absolute values and using the Boundedness property gives the desired claim. ■

Proof of Claim 3.4. Suppose that $h \in \mathrm{Lin}(\mathcal{C}, W)$ of the form $h(x) = \sum_{c\in\mathcal{C}} w_c \cdot c(x)$. From Claim 2.7, we know that the multicalibration violation for $c \in \mathcal{C}$ is bounded by $\alpha(v)$ for every $v \in \mathrm{Im}(\tilde{p})$.

$$\begin{aligned}|\mathbf{E}[h(\mathbf{x})(\mathbf{y}-v) \mid \tilde{p}(\mathbf{x}) = v]| &= \left|\mathbf{E}\left[\sum_{c\in\mathcal{C}} w_c \cdot c(\mathbf{x})(\mathbf{y}-v) \mid \tilde{p}(\mathbf{x}) = v\right]\right| \\ &\le \left(\sum_{c\in\mathcal{C}} |w_c|\right) \cdot \max_{c\in\mathcal{C}} |\mathbf{E}[c(\mathbf{x})(\mathbf{y}-v) \mid \tilde{p}(\mathbf{x}) = v]| \\ &\le W \cdot \alpha(v)\end{aligned}$$

The inequalities follow by Holder's inequality and the assumed bound on the weight of W for $h \in \mathrm{Lin}(\mathcal{C}, W)$. ■

Proof of Claim 3.5. Recall that $\mathrm{Cov}[\mathbf{y}, \mathbf{z}] = \mathbf{E}[\mathbf{yz}] - \mathbf{E}[\mathbf{y}]\,\mathbf{E}[\mathbf{z}]$. For any $h \in \mathrm{Lin}(\mathcal{C}, W)$ we have

$$\begin{aligned}|\mathrm{Cov}[\mathbf{y}, h(\mathbf{x})|\tilde{p}(\mathbf{x}) = v]| &= |\mathbf{E}[h(\mathbf{x})(\mathbf{y} - \mathbf{E}[\mathbf{y}])|\tilde{p}(\mathbf{x}) = v]| \\ &= |\mathbf{E}[h(\mathbf{x})(\mathbf{y}-v)|\tilde{p}(x) = v]| + |\mathbf{E}[(v-\mathbf{y})|\tilde{p}(\mathbf{x}) = v]| \\ &\le (W+1)\alpha(v)\end{aligned}$$

where we use the fact that $h \in \mathrm{Lin}(\mathcal{C}, W)$ and $1 \in \mathcal{C}$. Since $\mathbf{y} \in \{0,1\}$, this implies the claimed bounds by standard properties of covariance (see [15, Corollary 5.1]). ■

Proof of Lemma 3.6. For any $y \in \{0,1\}$,

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{D}|_v}[\ell(\mathbf{y}, h(\mathbf{x}))|(\tilde{p}(\mathbf{x}), \mathbf{y}) = (v, y)] &= \mathop{\mathbf{E}}_{\mathcal{D}|_v}[\ell(y, h(\mathbf{x}))|(\tilde{p}(\mathbf{x}), \mathbf{y}) = (v, y)] \\ &\ge \ell(y, \mathbf{E}[h(\mathbf{x})|(\tilde{p}(\mathbf{x}), \mathbf{y}) = (v, y)]) \qquad (16) \\ &= \ell(y, \mu(h : v, y)) \\ &\ge \ell(y, \Pi_\ell(\mu(h : v, y))). \qquad (17)\end{aligned}$$

where Equation (16) uses Jensen's inequality, and Equation (17) uses the optimality of projection for nice loss functions. Further, by the 1-Lipschitzness of ℓ on I_ℓ, and of Π_ℓ on $\mathbb{R}$

$$\begin{aligned}\ell(y, \Pi_\ell(\mu(h:v,y))) - \ell(y, \Pi_\ell(\mu(h:v))) &\le |\Pi_\ell(\mu(h:v,y)) - \Pi_\ell(\mu(h:v))| \\ &\le |\mu(h:v,y) - \mu(h:v)| \end{aligned} \tag{18}$$

Hence we have

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathcal{D}|_v}[\ell(\mathbf{y}, \Pi_\ell(\mu(h:v)))] - \mathop{\mathbf{E}}_{\mathcal{D}|_v}[\ell(\mathbf{y}, h(\mathbf{x}))] \\ &= \sum_{y\in\{0,1\}} \Pr[\mathbf{y}=y|\tilde{p}(\mathbf{x})=v]\left(\ell(y,\Pi_\ell(\mu(h:v))) - \mathbf{E}[\ell(y,h(\mathbf{x}))|(\tilde{p}(\mathbf{x}),\mathbf{y})=(v,y)]\right) \\ &\le \sum_{y\in\{0,1\}} \Pr[\mathbf{y}=y|\tilde{p}(\mathbf{x})=v]\left(\ell(y,\Pi_\ell(\mu(h:v))) - \ell(y,\Pi_\ell(\mu(h:v,y)))\right) \quad \text{(By Equation (17))} \\ &\le \sum_{y\in\{0,1\}} \Pr[\mathbf{y}=y|\tilde{p}(\mathbf{x})=v]\,|\mu(h:v,y) - \mu(h:v)| \quad \text{(by Equation (18))} \\ &\le 2(W+1)\alpha(v). \quad \text{(By Equation (9))}\end{aligned}$$

■

B Details on Algorithm

Here, we give a high-level overview of the MCBoost algorithm of [20] and weak agnostic learning.

Definition B.1 (Weak agnostic learning)**.** *Suppose $\mathcal{D}$ is a data distribution supported on $\mathcal{X}\times[-1,1]$. For a hypothesis class $\mathcal{C}$, a weak agnostic learner* WAL *solves the following promise problem: for some accuracy parameter $\alpha > 0$, if there exists some $c \in \mathcal{C}$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}}[c(\mathbf{x})\cdot\mathbf{z}] \ge \alpha$$

then WAL_α returns some $h : \mathcal{X} \to \mathbb{R}$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}}[h(\mathbf{x})\cdot\mathbf{z}] \ge \mathrm{poly}(\alpha).$$

For the sake of this presentation, we are informal about the polynomial factor in the guarantee of the weak agnostic learner. The smaller the exponent, the stronger the learning guarantee (i.e., we want WAL_α to return a hypothesis with correlation with $\mathbf{z}$ as close to $\Omega(\alpha)$ as possible). Standard arguments based on VC-dimension demonstrate that weak agnostic learning is statistically efficient.

B.1 MCBoost

The work introducing multicalibration [20] gives a boosting-style algorithm for learning multicalibrated predictors that has come to be known as MCBoost. The algorithm is an iterative procedure: starting with a trivial predictor, the MCBoost searches for a supported value $v \in \mathrm{Im}(\tilde{p})$ and "subgroup" $c_v \in \mathcal{C}$ that violate the multicalibration condition. Note that some care has to be taken to ensure that the predictor $\tilde{p}$ stays supported on finitely many values, and that each of these values

Algorithm 2 MCBoost

Parameters: hypothesis class $\mathcal{C}$ and $\alpha > 0$
Given: Dataset S sampled from $\mathcal{D}$
Initialize: $\tilde{p}(x) \leftarrow 1/2$.
Repeat:
if $\exists v \in \mathrm{Im}(\tilde{p})$ and $c_v \in \mathcal{C}$ such that

$$\mathbf{E}[c_v(\mathbf{x})\cdot(\mathbf{y}-v) \mid \tilde{p}(\mathbf{x}) = v] > \mathrm{poly}(\alpha) \tag{19}$$

update $\tilde{p}(x) \leftarrow \tilde{p}(x) + \eta c_v(x)\cdot\mathbf{1}[\tilde{p}(x)=v]$
Return: $\tilde{p}$

maintains significant measure in the data distribution $\mathcal{D}_{\tilde{p}}$. In this pseudocode, we ignore these issues; [20] handles them in full detail.

Importantly, the search over $\mathcal{C}$ for condition (19) can be reduced to weak agnostic learning. Intuitively, we pass WAL samples drawn from the data distribution, but labeled according to $\mathbf{z} = \mathbf{y} - v$ when $\tilde{p}(\mathbf{x}) = v$.

Lemma 3.8. The iteration complexity of MCBoost is directly (inverse quadratically) related to the size of the multicalibration violations we discover in (19). A standard potential argument can be found in [20].

By the termination condition, we can see that $\tilde{p}$ must actually be $(\mathcal{C}, \alpha)$-swap multicalibrated. In particular, when the algorithm terminates, then for all $v \in \mathrm{Im}(\tilde{p})$, we have that

$$\max_{c_v \in \mathcal{C}} \mathbf{E}[c_v(\mathbf{x}) \cdot (\mathbf{y} - v) \mid \tilde{p}(\mathbf{x}) = v] \leq \mathrm{poly}(\alpha) \leq \alpha.$$

Therefore, averaging over $\mathbf{v} \sim \mathcal{D}_{\tilde{p}}$, we obtain the guarantee. ■

Corollary 3.9. By Lemma 3.8, we know that $\tilde{p}$ returned by MCBoost is $(\mathcal{C}, \alpha)$-swap multicalibrated. By Theorem 3.3, $\tilde{p}$ is equivalently a $(\mathcal{L}_{\mathrm{cvx}}, \mathcal{C}, \alpha')$-swap omnipredictor for some polynomially-related α'. In other words, by Claim 2.4, if we post-process $\tilde{p}$ according to k_ℓ for any nice convex loss function ℓ, we obtain an $(\ell, \mathcal{C}, \varepsilon)$-swap agnostic learner. Taking $\alpha = \mathrm{poly}(\varepsilon)$ sufficiently small, we obtain the swap agnostic learning guarantee. ■

C Swap Loss Outcome Indistinguishability

In this Appendix, we give a full account of the definitions and results stated in Section 4. We introduce a unified notion of Swap Loss Outcome Indistinguishability, which captures all of the other notions of mutlicalibration and omniprediction defined so far. The notion builds on a line of work due to [6, 7], which propose the notion of *Outcome Indistinguishability* (OI) as a solution concept for supervised learning based on computational indistinguishability. In fact, the main result of [6] is an equivalence between OI and multicalibration. Despite the fact that OI is really multicalibration in disguise, the perspective has proved to be a useful technical perspective.

Key to this section is the prior work of [14]. This work proposes a new variant of OI, called *Loss OI*. The main result of [14] derives novel omniprediction guarantees from loss OI. Further, they show how to achieve loss OI using only calibration and multiaccuracy over a class of functions derived from the loss class $\mathcal{L}$ and hypothesis class $\mathcal{C}$. As we'll see, this class plays a role in the study of swap loss OI: swap loss OI is equivalent to multicalibration over the augmented class.

Additional Preliminaries. Intuitively, OI requires that outcomes sampled from the predictive model $\tilde{p}$ are indistinguishable from Nature's outcomes. Formally, we use $(\mathbf{x}, \mathbf{y}^*)$ to denote a sample from the true joint distribution over $\mathcal{X} \times \{0,1\}$. Then, given a predictor $\tilde{p}$, we associate it with the random variable with $\mathbf{E}[\tilde{\mathbf{y}}|x] = \tilde{p}(x)$, i.e., where $\tilde{\mathbf{y}}|x \sim \mathrm{Ber}(\tilde{p}(x))$. The variable $\tilde{\mathbf{y}}$ can be viewed as $\tilde{p}$'s simulation of Nature's label $\mathbf{y}^*$. In this section, we use $\mathcal{D}$ to denote the joint distribution $(\mathbf{x}, \mathbf{y}^*, \tilde{\mathbf{y}})$, where $\mathbf{E}[\mathbf{y}^*|x] = p^*(x)$ and $\mathbf{E}[\tilde{\mathbf{y}}|x] = \tilde{p}(x)$. While the joint distribution of $(\mathbf{y}^*, \tilde{\mathbf{y}})$ is not important to us, for simplicity we assume they are independent given $\mathbf{x} = x$.

C.1 Swap Loss OI

The notion of loss outcome indistinguishability was introduced in the recent work of [14] with the motivation of understanding omniprediction from the perspective of outcome indistinguishability [6]. Loss OI gives a strengthening of omniprediction. It requires predictors $\tilde{p}$ to fool a family $\mathcal{U}$ of statistical tests $u : \mathcal{X} \times [0,1] \times \{0,1\}$ that take a point $\mathbf{x} \in \mathcal{X}$, a prediction $\tilde{p}(\mathbf{x}) \in [0,1]$ and a label $\mathbf{y} \in \{0,1\}$ as their arguments. The goal is distinguish between the scenarios where $\mathbf{y} = \mathbf{y}^*$ is generated by *nature* versus where $\mathbf{y} = \tilde{\mathbf{y}}$ is a simulation of nature according to the predictor $\tilde{p}$. Formally, we require than for every $u \in \mathcal{U}$,

$$\mathop{\mathbf{E}}_{\mathcal{D}}[u(\mathbf{x}, \tilde{p}(\mathbf{x}), \mathbf{y}^*)] \approx_\varepsilon \mathop{\mathbf{E}}_{\mathcal{D}}[u(\mathbf{x}, \tilde{p}(\mathbf{x}), \tilde{\mathbf{y}})].$$

Loss OI specializes this to a specific family of tests arising in the analysis of omnipredictors.

Definition C.1 (Loss OI, [14])**.** *For a collection of loss functions $\mathcal{L}$, hypothesis class $\mathcal{C}$, and $\varepsilon \geq 0$, define the family of tests $\mathcal{U}(\mathcal{L}, \mathcal{C}) = \{u_{\ell,c}\}_{\ell \in \mathcal{L}, c \in \mathcal{C}}$ where*

$$u_{\ell,c}(x, v, y) = \ell(y, k_\ell(v)) - \ell(y, c(x)). \tag{20}$$

A predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ is $(\mathcal{L}, \mathcal{C}, \varepsilon)$-loss OI if for every $u \in \mathcal{U}(\mathcal{L}, \mathcal{C})$, it holds that

$$\left| \mathop{\mathbf{E}}_{(\mathbf{x}, \mathbf{y}^*) \sim \mathcal{D}} [u(\mathbf{x}, \tilde{p}(\mathbf{x}), \mathbf{y}^*)] - \mathop{\mathbf{E}}_{(\mathbf{x}, \tilde{\mathbf{y}}) \sim \mathcal{D}(\tilde{p})} [u(\mathbf{x}, \tilde{p}(\mathbf{x}), \tilde{\mathbf{y}})] \right| \leq \varepsilon. \tag{21}$$

[14] show that loss-OI implies omniprediction.

Lemma C.2 (Proposition 4.5, [14])**.** *If the predictor $\tilde{p}$ is $(\mathcal{L}, \mathcal{C}, \varepsilon)$-loss OI, then it is an $(\mathcal{L}, \mathcal{C}, \varepsilon)$-omnipredictor.*

Indeed, if the expected value of u is nonpositive for all $u \in \mathcal{U}(\mathcal{L}, \mathcal{C})$, then $\tilde{p}$ must achieve loss competitive with all $c \in \mathcal{C}$. The argument leverages the fact that u must be nonpositive when $\tilde{\mathbf{y}} \sim \mathrm{Ber}(\tilde{p}(\mathbf{x}))$—after all, in this world $\tilde{p}$ is the Bayes optimal. By indistinguishability, $\tilde{p}$ must also be optimal in the world where outcomes are drawn as $\mathbf{y}^*$. The converse, however, is not always true.

Next, we introduce swap loss OI, which allows the choice of distinguisher to depend on the predicted value.

Definition C.3 (Swap Loss OI)**.** *For a collection of loss functions $\mathcal{L}$, hypothesis class $\mathcal{C}$ and $\varepsilon \geq 0$, for an assignment of loss functions $\{\ell_v \in \mathcal{L}\}_{v \in \mathrm{Im}(\tilde{p})}$ and hypotheses $\{h_v \in \mathcal{H}\}_{v \in \mathrm{Im}(\tilde{p})}$, denote $u_v = u_{\ell_v, c_v} \in \mathcal{U}(\mathcal{L}, \mathcal{C})$. A predictor $\tilde{p}$ is $(\mathcal{L}, \mathcal{C}, \alpha)$-swap loss OI if for all such assignments,*

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [u_{\mathbf{v}}(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) - u_{\mathbf{v}}(\mathbf{x}, \mathbf{v}, \tilde{\mathbf{y}})] \right| \leq \alpha.$$

The notion generalizes both swap omniprediction and loss-OI simultaneously.

Lemma C.4. *If the predictor $\tilde{p}$ satisfies $(\mathcal{L}, \mathcal{C}, \alpha)$-swap loss OI, then*

- *it is an $(\mathcal{L}, \mathcal{C}, \alpha)$-swap omnipredictor.*
- *it is $(\mathcal{L}, \mathcal{C}, \alpha)$-loss OI.*

Proof. The proof of Part (1) follows the proof of [14, Proposition 4.5], showing that loss OI implies omniprediction. By the definition of k_{ℓ_v}, for every $x \in \mathcal{X}$ such that $\tilde{p}(x) = v$

$$\begin{aligned}
\mathop{\mathbf{E}}_{\tilde{\mathbf{y}} \sim \mathrm{Ber}(v)} u_v(x, v, \tilde{\mathbf{y}}) &= \mathop{\mathbf{E}}_{\tilde{\mathbf{y}} \sim \mathrm{Ber}(v)} [\ell_v(\tilde{\mathbf{y}}, k_{\ell_v}(v)) - \ell_v(\tilde{\mathbf{y}}, c_v(x))] \\
&= \ell_v(v, k_{\ell_v}(v)) - \ell_v(v, c_v(x)) \\
&\leq 0
\end{aligned}$$

Hence this also holds in expectation under $\mathcal{D}|_v$, which only considers points where $\tilde{p}(\mathbf{x}) = v$:

$$\mathop{\mathbf{E}}_{\mathcal{D}|_v} [u_v(\mathbf{x}, v, \tilde{\mathbf{y}})] \leq 0.$$

Since $\tilde{p}$ satisfies swap loss OI, we deduce that

$$\mathop{\mathbf{E}}_{\mathcal{D}|_v} [u_{\mathbf{v}}(\mathbf{x}, v, \mathbf{y}^*)] \leq \alpha(v)$$

Taking expectations over $\mathbf{v} \sim \mathcal{D}_{\tilde{p}}$ and using the definition of u_v, we get

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell_{\mathbf{v}}(\mathbf{y}^*, k_{\ell_{\mathbf{v}}}(\mathbf{v})) - \ell_{\mathbf{v}}(\mathbf{y}^*, c_{\mathbf{v}}(\mathbf{x}))] &= \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \mathop{\mathbf{E}}_{\mathcal{D}} [u_{\mathbf{v}}(\mathbf{x}, \mathbf{v}, \mathbf{y}^*)] \\
&\leq \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} [\alpha(\mathbf{v})] \leq \alpha
\end{aligned}$$

Rearranging the outer inequality gives

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell_{\mathbf{v}}(\tilde{\mathbf{y}}, k_{\ell_{\mathbf{v}}}(\mathbf{v}))] \leq \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell_{\mathbf{v}}(\mathbf{y}^*, c_{\mathbf{v}}(\mathbf{x}))] + \alpha.$$

Part (2) is implied by taking $\ell_v = \ell$ for every v. ∎

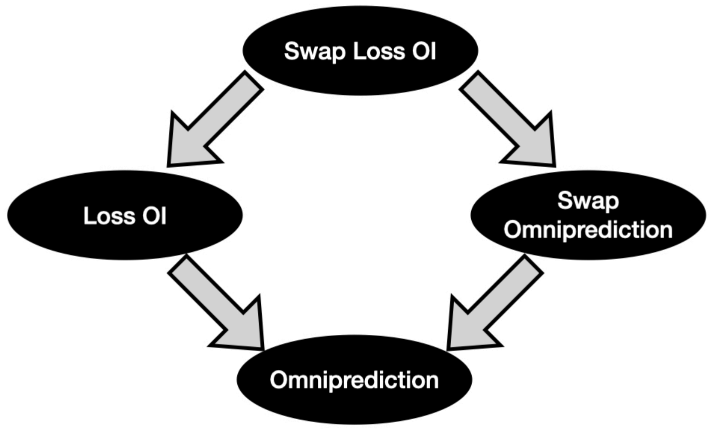

Figure 1: Relation between notions of omniprediction

C.2 Relating notions of omniprediction

In this work, we have discussed the four different notions of omniprediction defined to date.

00) Omniprediction, as originally defined by [15].

01) Loss OI, from [14].

10) Swap omniprediction.

11) Swap Loss OI.

In order to compare them, we can ask which of these notions implies the other for any fixed choice of loss class $\mathcal{L}$ and hypothesis class $\mathcal{C}$.

- Loss OI implies omniprediction by [14, Proposition 4.5].
- Swap omniprediction implies omniprediction by Claim 2.4.
- Swap loss OI implies both loss OI and swap multicalibration by Lemma C.4.

These relationships are summarized in Figure 1.

Further, this picture captures all the implications that hold for all $(\mathcal{L}, \mathcal{C})$. Next, we show that for any implication not drawn in the diagram, there exists some (natural) choice of $(\mathcal{L}, \mathcal{C})$, where the implication does not hold. In particular, we prove Theorem 4.2 which states that neither loss OI nor swap omniprediction implies the other for all $(\mathcal{L}, \mathcal{C})$. This separates these notions from swap loss OI, since swap loss OI implies both these notions.[5] By similar reasoning, it separates omniprediction from both these loss OI and swap omniprediciton, since omniprediction is implied by either of them.

Swap omniprediction does not imply loss OI. We prove this non-implication using a counterexample used in [14]. In particular, they show that omniprediction does not imply loss OI [14, Theorem 4.6], and the same example in fact shows that swap omniprediction does not imply loss OI. In their example, we have $\mathcal{D}$ on $\{\pm 1\}^3 \times [0, 1]$ where the marginal on $\{\pm 1\}^3$ is uniform, and $p^*(x) = (1 + x_1 x_2 x_3)/2$, whereas $\tilde{p}(x) = 1/2$ for all x. We take $\mathcal{C} = \{1, x_1, x_2, x_3\}$. Since $\tilde{p} = 1/2$ is constant, it is easy to check that $\tilde{p} - p^* = -x_1 x_2 x_3/2$ is uncorrelated with $\mathcal{C}$. Hence $\tilde{p}$ satisfies swap multicalibration (which is the same as multicalibration or even multiaccuracy in this setting where $\tilde{p}$ is constant). Hence by Theorem 3.3, $\tilde{p}$ is an $(\mathcal{L}_{\mathsf{cvx}}(1), \mathrm{Lin}_{\mathcal{C}}, 0)$-swap omnipredictor. [14, Theorem 4.6] prove that $\tilde{p}$ is not loss OI for the ℓ_4 loss. Hence we have the following result.

[5] For instance if loss OI implied swap loss OI, it would also imply swap omniprediction, which our claim shows it does not.

$x = (x_1, x_2)$	$p^*(x)$	$\tilde{p}(x)$
$(-1, -1)$	0	$\frac{1}{8}$
$(+1, -1)$	$\frac{1}{4}$	$\frac{1}{8}$
$(-1, +1)$	1	$\frac{7}{8}$
$(+1, +1)$	$\frac{3}{4}$	$\frac{7}{8}$

Table 2: Separating loss-OI and swap-resilient omniprediction

Lemma C.5. *The predictor $\tilde{p}$ is $(\mathcal{C}, 0)$-swap multicalibrated and hence it is a $(\{\ell_4\}, \mathrm{Lin}(\mathcal{C}), 0)$-swap omnipredictor. But it is not $(\{\ell_4\}, \mathrm{Lin}(\mathcal{C}, 1), \varepsilon)$-loss OI for $\varepsilon < 4/9$.*

We remark that the construction extends to all ℓ_p losses for even $p > 2$. Hence even for convex losses, the notions of swap omniprediction are loss-OI seem incomparable.

Loss OI does not imply swap omniprediction. Next we construct an example showing that loss OI need not imply swap omniprediction. We consider the set of all GLM losses defined below, which contain common losses including the squared loss and the logistic loss.

Definition C.6. *Let $g : \mathbb{R} \to \mathbb{R}$ be a convex, differentiable function such that $[0, 1] \subseteq \mathrm{Im}(g')$. Define its matching loss to be $\ell_g = g(t) - yt$. Define $\mathcal{L}_{\mathsf{GLM}} = \{\ell_g\}$ be the set of all such loss functions.*

[14] shows a general decomposition result that reduces achieving loss OI to a calibration condition and a multiaccuracy condition. Whereas arbitrary losses might require multiaccuracy for the more powerful class $\partial\mathcal{L} \circ \mathcal{C}$, for $\mathcal{L}_{\mathsf{GLM}}$, $\partial\mathcal{L}_{\mathsf{GLM}} \circ \mathcal{C} = \mathcal{C}$. This is formalized in the following result.

Lemma C.7 (Theorem 5.3, [14])**.** *If $\tilde{p}$ is ε_1-calibrated and $(\mathcal{C}, \varepsilon_2)$-multiaccurate, then it is $(\mathcal{L}_{\mathsf{GLM}}, \mathrm{Lin}(\mathcal{C}, W), \delta)$-loss OI for $\delta = \varepsilon_1 + W\varepsilon_2$.*

In light of the above result, it suffices to find a predictor that is calibrated and multiaccurate (and hence satisfies loss OI), but not multicalibrated, hence not swap multicalibrated. By Theorem 3.3 it is not an $(\{\ell_2\}, \mathrm{Lin}_{\mathcal{C}}, \delta)$-swap omnipredictor for δ less than some constant.

Let us define the predictors $p^*, \tilde{p} : \{\pm 1\}^2 \to [0, 1]$ as below. We use these to show a separation between loss OI and swap omniprediction.

Lemma C.8. *Consider the distribution $\mathcal{D}$ on $\{\pm 1\}^2 \times \{0, 1\}$ where the marginal on $\{\pm 1\}^2$ is uniform and $\mathbf{E}[\mathbf{y}|x] = p^*(x)$. Let $\mathcal{C} = \{1, x_1, x_2\}$.*

1. *$\tilde{p} \in \mathrm{Lin}(\mathcal{C}, 1)$. Moreover, it minimizes the squared error over all hypotheses from $\mathrm{Lin}(\mathcal{C})$.*
2. *$\tilde{p}$ is perfectly calibrated and $(\mathcal{C}, 0)$-multiaccurate. So it is $(\mathcal{L}_{\mathsf{GLM}}, \mathrm{Lin}(\mathcal{C}), 0)$-loss OI.*
3. *$\tilde{p}$ is not $(\mathcal{C}, \alpha)$-multicalibrated for $\alpha < 1/8$. It is not $(\ell_2, \mathrm{Lin}(\mathcal{C}), \delta)$-swap agnostic learner for $\delta < 1/64$.*

Proof. We compute Fourier expansions for the two predictors:

$$p^*(x) = \frac{1}{8}(4 + 3x_2 - x_1 x_2) \tag{22}$$

$$\tilde{p}(x) = \frac{1}{8}(4 + 3x_2) \tag{23}$$

This shows that $\tilde{p} \in \mathrm{Lin}(\mathcal{C})$, and moreover that it is the optimal approximation to p^* in $\mathrm{Lin}(\mathcal{C})$, as it is the projection of p^* onto $\mathrm{Lin}(\mathcal{C})$. This shows that $\tilde{p}$ is an $(\ell_2, \mathrm{Lin}(\mathcal{C}), 0)$-agnostic learner.

It is easy to check that $\tilde{p}$ is perfectly calibrated. It is $(\mathcal{C}, 0)$-multiaccurate, since it is the projection of p^* onto $\mathrm{Lin}(\mathcal{C})$, so $\tilde{p} - p^*$ is orthogonal to $\mathrm{Lin}(\mathcal{C})$. Hence we can apply Lemma C.7 to conclude that it is $(\mathcal{L}_{\mathsf{GLM}}, \mathrm{Lin}(C), 0)$-loss OI, where $\mathcal{L}_{\mathsf{GLM}}$ which contains the squared loss.

To show that $\tilde{p}$ is not swap-agnostic, we observe that conditioning on the value of $\tilde{p}(\mathbf{x}) = (4 + 3x_2)/8$ is equivalent to conditioning on $x_2 \in \{\pm 1\}$. For each value of x_2, the restriction of p^* which is now linear in x_1 belongs to $\mathrm{Lin}(\mathcal{C})$. Indeed if we condition on $\tilde{p}(x) = 1/8$ so that $x_2 = -1$, we have

$$p^*(x) = \frac{1}{2} - \frac{3}{8} + \frac{1}{8}x_1 = \frac{1 + x_1}{8}.$$

Conditioned on $\tilde{p}(x) = 7/8$ so that $x_2 = 1$, we have

$$p^*(x) = \frac{1}{2} + \frac{3}{8} - \frac{1}{8}x_1 = \frac{7 - x_1}{8}.$$

Hence we have

$$\mathop{\mathbf{E}}_{v \sim \mathcal{D}_{\tilde{p}}}\left[\left|\min_{h \in \mathrm{Lin}(\mathcal{C})} \mathbf{E}[(\mathbf{y} - h(\mathbf{x}))^2 | f(\mathbf{x}) = v]\right|\right] = \mathbf{E}[(y - p^*(x))^2] = \mathrm{Var}[\mathbf{y}],$$

whereas the variance decomposition of squared loss gives

$$\begin{aligned}\mathbf{E}[(\mathbf{y} - \tilde{p}(\mathbf{x}))^2] &= \mathbf{E}[(\mathbf{y} - p^*(\mathbf{x}))^2] + \mathbf{E}[(p^*(\mathbf{x}) - \tilde{p}(\mathbf{x}))^2] \\ &= \mathrm{Var}[\mathbf{y}] + \frac{1}{64}\mathbf{E}[(x_1 x_2)^2] \\ &= \mathrm{Var}[\mathbf{y}] + \frac{1}{64}.\end{aligned}$$

Hence $\tilde{p}$ is not a $(\ell_2, \mathrm{Lin}(\mathcal{C}), \delta)$-swap agnostic learner for $\delta < 1/64$.

To see that f is not multicalibrated for small α, observe that conditioned on $x_2 \in \{\pm 1\}$, the correlation between x_1 and $\tilde{p} - p^*$ is $1/8$. ∎

Note that item (1) above separates swap omniprediction from omniprediction and agnostic learning. This separation can also be derived from [15, Theorem 7.5] which separated (standard) omniprediction from agnostic learning, since swap omniprediction implies standard omniprediction.

Comparing notions for GLM losses. When we restrict our attention to $\mathcal{L}_{\mathsf{GLM}}$, in fact, the notions of swap loss OI and swap omniprediction are equivalent. The key observation here is that $\partial\mathcal{L}_{\mathsf{GLM}} \circ \mathcal{C} = \mathcal{C}$, as shown in [14]. Paired with Theorem 3.3 and Theorem 4.1 (proved next), we obtain the following collapse.

Claim C.9. *The notions of $(\mathcal{L}_{\mathsf{GLM}}, \mathcal{C}, \alpha_1)$-swap loss OI and $(\mathcal{L}_{\mathsf{GLM}}, \mathcal{C}, \alpha_2)$-swap omniprediction are equivalent.*

Proof. To see this, note that by Theorem 4.1, $(\mathcal{L}_{\mathsf{GLM}}, \mathcal{C}, \alpha_1)$-swap loss OI is equivalent to $(\partial\mathcal{L}_{\mathsf{GLM}} \circ \mathcal{C}, \alpha_1')$-swap multicalibration. We know from Theorem 3.3 that this is also equivalent to $(\partial\mathcal{L}_{\mathsf{GLM}} \circ \mathcal{C}, \alpha_2)$-swap omniprediction. So, by the fact that $\partial\mathcal{L}_{\mathsf{GLM}} \circ \mathcal{C} = \mathcal{C}$, we have the claimed equivalence. ∎

Finally, we know that loss OI implies omniprediction for $\mathcal{L}_{\mathsf{GLM}}$, since this holds true for all $\mathcal{L}$. We do not know if these notions are equivalent for $\mathcal{L}_{\mathsf{GLM}}$, since the construction in Lemma C.5 used the ℓ_4 loss which does not belong to $\mathcal{L}_{\mathsf{GLM}}$.

C.3 Equivalence of swap loss OI and swap multicalibration over augmented class

We show that $(\mathcal{L}, \mathcal{C})$-swap loss OI and $(\partial\mathcal{L} \circ \mathcal{C})$-swap multicalibration are equivalent for nice loss functions.

Theorem C.10 (Formal statement of Theorem 4.1)**.** *Let $\mathcal{L} \subseteq \mathcal{L}(B)$ be a family of B-nice loss functions such that $\ell_2 \in \mathcal{L}$. Then $(\partial\mathcal{L} \circ \mathcal{C}, \alpha_1)$-swap multicalibration and $(\mathcal{L}, \mathcal{C}, \alpha_2)$-swap loss OI are equivalent.*[6]

In preparation for this, we start with the following simple claim from [14].

Claim C.11 (Lemma 4.8, [14])**.** *For random variables $\mathbf{y}_1, \mathbf{y}_2 \in \{0, 1\}$ and $t \in \mathbb{R}$,*

$$\mathbf{E}[\ell(\mathbf{y}_1, t) - \ell(\mathbf{y}_2, t)] = \mathbf{E}[(\mathbf{y}_1 - \mathbf{y}_2)\partial\ell(t)]. \tag{24}$$

[6] Here equivalence means that there are reductions in either direction that lose a multiplicative factor of $(B + 1)$ in the error.

We record two corollaries of this claim. These can respectively be seen as strengthenings of the two parts of Theorem [14, Theorem 4.9], which respectively characterized hypothesis OI in terms of multiaccuracy and decision OI in terms of calibration. We generalize these to the swap setting.

Corollary C.12. *For every choice of* $\{\ell_v, c_v\}_{v \in \mathrm{Im}(\tilde{p})}$*, we have*

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell_{\mathbf{v}}(\mathbf{y}^*, c_{\mathbf{v}}(\mathbf{x})) - \ell_{\mathbf{v}}(\tilde{\mathbf{y}}, c_{\mathbf{v}}(\mathbf{x}))] \right| \right] = \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [(\mathbf{y}^* - \tilde{\mathbf{y}}) \partial \ell_{\mathbf{v}} \circ c_{\mathbf{v}}(\mathbf{x})] \right| \right]. \tag{25}$$

Hence if $\tilde{p}$ *is* $(\partial \mathcal{L} \circ \mathcal{C}, \alpha)$*-swap multicalibrated, then*

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell_{\mathbf{v}}(\mathbf{y}^*, c_{\mathbf{v}}(\mathbf{x})) - \ell_{\mathbf{v}}(\tilde{\mathbf{y}}, c_{\mathbf{v}}(\mathbf{x}))] \right| \right] \leq \alpha.$$

Proof. Equation (25) is derived by applying Equation (24) to the LHS. Assuming that $\tilde{p}$ is $(\partial \mathcal{L} \circ \mathcal{C}, \alpha)$-swap multicalibrated, we have

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [(\mathbf{y}^* - \tilde{\mathbf{y}}) \partial \ell_{\mathbf{v}} \circ c_{\mathbf{v}}(\mathbf{x})] \right| \right] \leq \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \max_{c' \in \partial \mathcal{L} \circ \mathcal{C}} \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [(\mathbf{y}^* - \tilde{\mathbf{y}}) c'(\mathbf{x})] \right| \right] \leq \alpha.$$

■

Corollary C.13. *Let* $\{\ell_v\}_{v \in \mathrm{Im}(f)}$ *be a collection of loss* B*-nice loss functions. Let* $k(v) = k_{\ell_v}(v)$*. If* $\tilde{p}$ *is* α*-calibrated then*

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell_{\mathbf{v}}(\mathbf{y}^*, k(\mathbf{v})) - \ell_{\mathbf{v}}(\tilde{\mathbf{y}}, k(\mathbf{v}))] \right| \right] \leq B\alpha. \tag{26}$$

Proof. We have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\ell_{\mathbf{v}}(\mathbf{y}^*, k(\mathbf{v})) - \ell_{\mathbf{v}}(\tilde{\mathbf{y}}, k(\mathbf{v}))] \right| \right] &= \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [(\mathbf{y}^* - \mathbf{v}) \partial \ell_{\mathbf{v}}(k(\mathbf{v}))] \right| \right] \\
&= \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[|\partial \ell_{\mathbf{v}}(k(\mathbf{v}))| \left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\mathbf{y}^* - \mathbf{v}] \right| \right] \\
&\leq B \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\mathbf{y}^* - \mathbf{v}] \right| \right] \\
&\leq B\alpha.
\end{aligned}$$

where we use the fact that $k(v) \in I_\ell$, and so $|\partial \ell_v(k(v))| \leq B$. ■

Finally, we show the following key technical lemma which explains why the ℓ_2 loss has a special role.

Lemma C.14. *If* $\tilde{p}$ *is* $(\{\ell_2\}, \mathcal{C}, \alpha)$*-swap OI, then it is* α*-calibrated.*

Proof. Observe that $\ell_2(y, v) = (y - v)^2/2$ so $k_{\ell_2}(v) = v$. Hence,

$$\begin{aligned}
u_{\ell_2, 0}(x, v, y) &= \ell_2(y, k_\ell(v)) - \ell_2(y, 0) \\
&= ((y - v)^2 - y^2)/2 \\
&= -vy + v^2/2.
\end{aligned} \tag{27}$$

Recall that $\{0, 1\} \subset \mathcal{C}$. The implication of swap loss OI when we take $c_v = 0$ for all v is that

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [u_{\ell_2, 0}(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) - u_{\ell_2, 0}(\mathbf{x}, \mathbf{v}, \tilde{\mathbf{y}})] \right| \right] \leq \alpha.$$

We can simplify the LHS using Equation (27) to derive

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [(-\mathbf{v}\mathbf{y}^* + \mathbf{v}^2/2) - (-\mathbf{v}\tilde{\mathbf{y}} + \mathbf{v}^2/2)] \right| \right] &= \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\mathbf{v}(\mathbf{y}^* - \tilde{\mathbf{y}})] \right| \right] \\
&= \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\mathbf{v} \left| \mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}} [\tilde{\mathbf{y}} - \mathbf{y}^*] \right| \right] \leq \alpha.
\end{aligned} \tag{28}$$

Considering the case where $c_v = 1$ for all v gives

$$\begin{aligned} u_{\ell_2,1}(x, v, y) &= \ell_2(y, k_\ell(v)) - \ell_2(y, 1) \\ &= ((y - v)^2 - (1 - y)^2)/2 \\ &= (1 - v)y + (v^2 - 1)/2. \end{aligned}$$

We derive the following implication of swap loss OI by taking $c_v = 0$ for all v:

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [u_{\ell_2,1}(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) - u_{\ell_2,1}(\mathbf{x}, \mathbf{v}, \tilde{\mathbf{y}})] \right| \right] = \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[(1 - \mathbf{v}) \left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\tilde{\mathbf{y}} - \mathbf{y}^*] \right| \right] \leq \alpha \tag{29}$$

Adding the bounds from Equations (28) and (29) we get

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\mathbf{v} - \mathbf{y}^*] \right| \right] = \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\tilde{\mathbf{y}} - \mathbf{y}^*] \right| \right] \leq \alpha$$

■

Proof of Theorem 4.1. We first show the forward implication, that swap multicalibration implies swap loss OI.

Since $\ell_2 \in \mathcal{L}$ and $1 \in \mathcal{C}$, we have $\partial \ell_2 \circ 1 = 1 \in \partial \mathcal{L} \circ \mathcal{C}$. This implies that $\tilde{p}$ is α-mulitcalibrated, since

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [1(\mathbf{y} - \mathbf{v})] \right| \right] \leq \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\max_{c \in \mathcal{C}} \left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [c(\mathbf{x})(\mathbf{y} - \mathbf{v})] \right| \right] \leq \alpha.$$

Consider any collection of losses $\{\ell_v\}_{v \in \mathrm{Im}(\tilde{p})}$. Applying Corollary C.13, we have

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\ell_\mathbf{v}(\mathbf{y}^*, k(\mathbf{v})) - \ell_\mathbf{v}(\tilde{\mathbf{y}}, k(\mathbf{v}))] \right| \right] \leq B\alpha.$$

On the other hand, by Corollary C.12, we have for every choice of $\{\ell_v, c_v\}_{v \in \mathrm{Im}(\tilde{p})}$,

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\ell_\mathbf{v}(\mathbf{y}^*, c_\mathbf{v}(\mathbf{x})) - \ell_\mathbf{v}(\tilde{\mathbf{y}}, c_\mathbf{v}(\mathbf{x}))] \right| \right] \leq \alpha.$$

Hence for any choice of $\{u_v\}_{v \in \mathrm{Im}(\tilde{p})}$ we can bound

$$\begin{aligned} &\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [u_\mathbf{v}(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) - u_\mathbf{v}(\mathbf{x}, \mathbf{v}, \tilde{\mathbf{y}})] \right| \\ &\leq \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\ell_\mathbf{v}(\mathbf{y}^*, k(\mathbf{v})) - \ell_\mathbf{v}(\tilde{\mathbf{y}}, k(\mathbf{v})] \right| + \left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\ell_\mathbf{v}(\mathbf{y}^*, c_\mathbf{v}(\mathbf{x})) - \ell_\mathbf{v}(\tilde{\mathbf{y}}, c_\mathbf{v}(\mathbf{x}))] \right| \right] \\ &\leq (B + 1)\alpha \end{aligned}$$

which shows that $\tilde{p}$ satisfies swap loss OI with $\alpha_2 = (D + 1)\alpha_1$.

Next we show the reverse implication: if $\tilde{p}$ satisfies $(\mathcal{L}, \mathcal{C}, \alpha_2)$-swap loss OI, then it satisfies $(\partial \mathcal{L} \circ \mathcal{C}, \alpha_1)$-swap multicalibration. The first step is to observe that by lemma C.14, since $\ell_2 \in \mathcal{L}$, the predictor $\tilde{p}$ is α_2 calibrated. Since any $\ell \in \mathcal{L}$ is B-nice, we have

$$\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\ell_\mathbf{v}(\mathbf{y}^*, k(\mathbf{v})) - \ell_\mathbf{v}(\tilde{\mathbf{y}}, k(\mathbf{v}))] \right| \right] = \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [(\mathbf{y}^* - \tilde{\mathbf{y}})k(\mathbf{v})] \right| \right] \leq B\alpha_2.$$

For any $\{\ell_v, c_v\}_{v \in \mathrm{Im}(f)}$, since

$$u_v(x, v, y) = \ell_v(y, k_\ell(v)) + \ell_v(y, c_v(x))$$

we can write

$$\begin{aligned} &\mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\ell_\mathbf{v}(\mathbf{y}^*, c(\mathbf{x})) - \ell_\mathbf{v}(\tilde{\mathbf{y}}, c(\mathbf{x}))] \right| \right] \\ &\leq \mathop{\mathbf{E}}_{\mathbf{v} \sim \mathcal{D}_{\tilde{p}}} \left[\left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [u_\mathbf{v}(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) - u_\mathbf{v}(\mathbf{x}, \mathbf{v}, \tilde{\mathbf{y}})] \right| + \left| \mathop{\mathbf{E}}_{\mathcal{D}|_\mathbf{v}} [\ell_\mathbf{v}(\mathbf{y}^*, k(\mathbf{v})) - \ell_\mathbf{v}(\tilde{\mathbf{y}}, k(\mathbf{v}))] \right| \right] \\ &\leq (B + 1)\alpha_2. \end{aligned}$$

But by Equation (25), the LHS can be written as

$$\mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}_{\tilde{p}}}\left[\left|\mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}}[\ell_{\mathbf{v}}(\mathbf{y}^*, c(\mathbf{x})) - \ell_{\mathbf{v}}(\tilde{\mathbf{y}}, c(\mathbf{x}))]\right|\right] = \mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}_{\tilde{p}}}\left[\left|\mathop{\mathbf{E}}_{\mathcal{D}|_{\mathbf{v}}}[\partial\ell_{\mathbf{v}} \circ c_{\mathbf{v}}(\mathbf{x})(\mathbf{y}^* - \mathbf{v})]\right|\right]$$

This shows that $\tilde{p}$ is $(\partial\mathcal{L} \circ \mathcal{C}, (B+1)\alpha_2)$-swap multicalibrated. ■