A Additional Material: Proofs for *Mixability in Statistical Learning*

Here we collect proofs that were omitted from the main body of the paper due to lack of space.

A.1 Proof of Proposition 1

Proof. As $\frac{e^{-\eta\ell(Y,f(X))}}{e^{-\eta\ell(Y,f^*(X))}} = e^{-\eta(\ell(Y,f(X))-\ell(Y,f^*(X)))}$ is convex in η, linearity of expectation implies that $\psi(\eta) := \mathbf{E}\left[\frac{e^{-\eta\ell(Y,f(X))}}{e^{-\eta\ell(Y,f^*(X))}}\right]$ is also convex in η. Observing that $\psi(0) = 1$, we have 0-stochastic mixability. And by $\psi(\gamma) = \psi\big((1-\frac{\gamma}{\eta})\cdot 0 + \frac{\gamma}{\eta}\cdot\eta\big) \leq (1-\frac{\gamma}{\eta})\psi(0) + \frac{\gamma}{\eta}\psi(\eta) \leq 1$ we obtain γ-stochastic mixability. $\square$

A.2 Proof of Theorem 2

Proof. Let f^* be as in Definition 2. For $\lambda \in [0,1]$ and any distribution π on $\mathcal{F}$, define the function

$$\phi_\pi(\lambda, x, y) = -\ln\left((1-\lambda)e^{-\eta\ell(y,f^*(x))} + \lambda\int e^{-\eta\ell(y,f(x))}\,\pi(\mathrm{d}f)\right), \tag{10}$$

and let $\phi_\pi(\lambda) = \mathbf{E}[\phi_\pi(\lambda, X, Y)]$ be its expectation. Then for any x and y, $\phi_\pi(\lambda, x, y)$ is convex in λ, because it is the composition of $-\ln$ with a linear function. By linearity of expectation, it follows that $\phi_\pi(\lambda)$ is also convex.

Stochastic mixability is related to $\phi'_\pi(0)$, the right-derivative of ϕ_π at $\lambda = 0$, which we will now compute. As $\phi_\pi(\lambda, x, y)$ is convex, the slope $s_\pi(h, x, y) = \frac{\phi_\pi(0+h,x,y)-\phi_\pi(0,x,y)}{h}$ is nondecreasing in h, and

$$s_\pi(1/2, x, y) = 2\ln\frac{e^{-\eta\ell(y,f^*(x))}}{\frac{1}{2}e^{-\eta\ell(y,f^*(x))} + \frac{1}{2}\int e^{-\eta\ell(y,f(x))}\,\pi(\mathrm{d}f)} \leq 2\ln\frac{e^{-\eta\ell(y,f^*(x))}}{\frac{1}{2}e^{-\eta\ell(y,f^*(x))}} = 2\ln 2.$$

Hence $\mathbf{E}[s_\pi(1/2, X, Y)] \leq 2\ln 2 < \infty$ and by the monotone convergence theorem [26]

$$\begin{aligned}\phi'_\pi(0) &= \lim_{h\downarrow 0}\mathbf{E}[s_\pi(h, X, Y)] = \mathbf{E}\big[\lim_{h\downarrow 0} s_\pi(h, X, Y)\big] = \mathbf{E}\big[\frac{\mathrm{d}}{\mathrm{d}\lambda}\phi_\pi(\lambda, X, Y)|_{\lambda=0}\big]\\ &= 1 - \mathbf{E}\big[\int \frac{e^{-\eta\ell(Y,f(X))}}{e^{-\eta\ell(Y,f^*(X))}}\,\pi(\mathrm{d}f)\big] = 1 - \int \mathbf{E}\big[\frac{e^{-\eta\ell(Y,f(X))}}{e^{-\eta\ell(Y,f^*(X))}}\big]\pi(\mathrm{d}f).\end{aligned}$$

Comparing to (3), we see that η-stochastic mixability is equivalent to the property that $\phi'_\pi(0) \geq 0$ for all π. And as ϕ_π is convex, this in turn is equivalent to $\phi_\pi(\lambda)$ being nondecreasing.

Suppose first that $(\ell, \mathcal{F}, P^*)$ is η-stochastically mixable. Then, for any π, $\phi_\pi(\lambda)$ is nondecreasing and hence

$$\eta\,\mathbf{E}[\ell(Y, f^*(X))] = \phi_\pi(0) \leq \phi_\pi(1) = \mathbf{E}\big[-\ln\int e^{-\eta\ell(Y,f(X))}\,\pi(\mathrm{d}f)\big],$$

from which (5) follows. Conversely, suppose that (5) holds for all π. Then it holds in particular for $\pi = (1-\lambda)\delta_{f^*} + \lambda\bar{\pi}$, where δ_{f^*} is a point-mass on f^*, $\lambda \in [0,1]$ is arbitrary, and $\bar{\pi}$ is an arbitrary distribution on $\mathcal{F}$. Plugging this choice of π into (5), we find that

$$\begin{aligned}&\frac{1}{\eta}\phi_{\bar{\pi}}(0) = \mathbf{E}[\ell(Y, f^*(X))]\\ &\qquad\leq \mathbf{E}\big[-\frac{1}{\eta}\ln\left((1-\lambda)e^{-\eta\ell(y,f^*(x))} + \lambda\int e^{-\eta\ell(y,f(x))}\,\bar{\pi}(\mathrm{d}f)\right)\big] = \frac{1}{\eta}\phi_{\bar{\pi}}(\lambda)\end{aligned}$$

for any λ and $\bar{\pi}$. It follows that $\phi_{\bar{\pi}}(\lambda)$ is minimized at $\lambda = 0$, and hence by its convexity that it is nondecreasing. As we have established that η-stochastic mixability is implied when $\phi_{\bar{\pi}}(\lambda)$ is nondecreasing for all $\bar{\pi}$, the proof is complete. $\square$

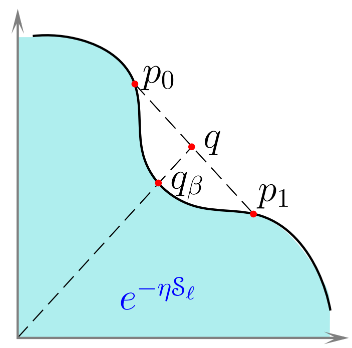

Figure 2: Illustration of the proof of Lemma 7.

A.3 Proof of Lemma 6

Proof. Let $f \in \mathcal{F}$ be arbitrary, and for $0 \le \lambda < 1$ define

$$\mu(\lambda) = \mathbf{E}\left[-\tfrac{1}{\eta}\ln\left((1-\lambda)e^{-\eta\ell(Y,f^*(X))} + \lambda e^{-\eta\ell(Y,f(X))}\right)\right].$$

Then η-mixability of ℓ implies that for any $x \in \mathcal{X}$ and λ there exists $a_\lambda(x) \in \mathcal{A}$ such that

$$\ell(y, a_\lambda(x)) \le -\tfrac{1}{\eta}\ln\left((1-\lambda)e^{-\eta\ell(y,f^*(x))} + \lambda e^{-\eta\ell(y,f(x))}\right) \qquad \forall y \in \mathcal{Y}.$$

Hence for any λ, we have $\mu(\lambda) \ge \mathbf{E}[\ell(Y, a_\lambda(X))] \ge \mathbf{E}[\ell(Y, f^*(X))] = \mu(0)$. This implies that $\mu'(0) \ge 0$, where $\mu'(\lambda)$ is the right-derivative of $\mu(\lambda)$, and the lemma follows by computing $\mu'(0)$:

$$\mu'(\lambda) = \tfrac{-1}{\eta}\,\mathbf{E}\left[\frac{e^{-\eta\ell(Y,f(X))} - e^{-\eta\ell(Y,f^*(X))}}{(1-\lambda)e^{-\eta\ell(Y,f^*(X))} + \lambda e^{-\eta\ell(Y,f(X))}}\right]$$

$$0 \le \eta\mu'(0) = \mathbf{E}\left[\frac{e^{-\eta\ell(Y,f^*(X))} - e^{-\eta\ell(Y,f(X))}}{e^{-\eta\ell(Y,f^*(X))}}\right] = 1 - \mathbf{E}\left[\frac{e^{-\eta\ell(Y,f(X))}}{e^{-\eta\ell(Y,f^*(X))}}\right]. \qquad \square$$

A.4 Proof of Lemma 7

Proof. Suppose that ℓ is not η-mixable. Then we will show that $(\ell, \mathcal{F}_{\text{full}})$ cannot be η-stochastically mixable either. Since ℓ is not η-mixable, there must exist $p_0, p_1 \in \Phi := e^{-\eta\mathcal{S}_\ell}$ and $\lambda \in (0,1)$ such that $q := (1-\lambda)p_0 + \lambda p_1$ is not in Φ (see Figure 2). For $i = 1, 2$, we have $-\frac{1}{\eta}\ln p_i \in \mathcal{S}_\ell$, so there must exist predictions $a_0, a_1 \in \mathcal{A}$ such that $\ell_{a_i}(y) \le -\frac{1}{\eta}\ln p_i(y)$ for all y or, equivalently, $e^{-\eta\ell_{a_i}(y)} \ge p_i(y)$. Let $f_i \in \mathcal{F}_{\text{full}}$ be such that $f_i(x) = a_i$ for all x. We will construct a distribution P^* on $\mathcal{X} \times \mathcal{Y}$ such that

$$\mathbf{E}_{P^*}\left[\ell\left(Y, f(X)\right)\right] > \mathbf{E}_{P^*}\left[-\tfrac{1}{\eta}\ln q(Y)\right] \tag{11}$$

for all $f \in \mathcal{F}_{\text{full}}$. But, by the monotonicity of $-\ln$, we have

$$\mathbf{E}_{P^*}\left[-\tfrac{1}{\eta}\ln q(Y)\right] \ge \mathbf{E}_{P^*}\left[-\tfrac{1}{\eta}\ln\left((1-\lambda)e^{-\eta\ell(Y,f_0(X))} + \lambda e^{-\eta\ell(Y,f_1(X))}\right)\right],$$

which contradicts η-stochastic mixability of $(\ell, \mathcal{F}_{\text{full}}, P^*)$ by the characterization in Theorem 2 for the distribution π that assigns point masses $1-\lambda$ and λ to f_0 and f_1, respectively.

Our approach to establish (11) is illustrated by Figure 2. We define $q_\alpha = \alpha q$ for $\alpha \in [0,1]$, and let $\beta = \sup\{\alpha \mid q_\alpha \in \Phi\}$. We will show that $\beta \in [0,1)$ and that q_β lies on the boundary of Φ. Then, by assumption, $-\frac{1}{\eta}\ln q_\beta$ is supportable, so that there exists a distribution P_Y^* on $\mathcal{Y}$ such that

$$\mathbf{E}_{P_Y^*}\left[-\tfrac{1}{\eta}\ln q_\beta(Y)\right] \le \mathbf{E}_{P_Y^*}[t(Y)] \qquad \text{for all } t \in \mathcal{S}. \tag{12}$$

Now let P_X^* be any distribution on $\mathcal{X}$ and define $P^* = P_X^* \times P_Y^*$. Then, for any $f \in \mathcal{F}_{\text{full}}$, (12) implies that

$$\begin{aligned}\mathbf{E}_{P^*}[\ell(Y, f(X))] &= \mathbf{E}_{P_X^*}\,\mathbf{E}_{P_Y^*}\left[\ell(Y, f(X)) \mid X\right] \\ &\geq \mathbf{E}_{P_X^*}\,\mathbf{E}_{P_Y^*}\left[-\tfrac{1}{\eta}\ln q_\beta(Y)\right] \\ &= \mathbf{E}_{P^*}\left[-\tfrac{1}{\eta}\ln q(Y)\right] - \tfrac{1}{\eta}\ln\beta \\ &> \mathbf{E}_{P^*}\left[-\tfrac{1}{\eta}\ln q(Y)\right],\end{aligned}$$

as required.

To show that $\beta \in [0,1)$, we first observe that $0 \leq q_0(y)$ for all y, so that $q_0 \in \Phi$ and hence $\beta \geq 0$. Furthermore, $q_\alpha \in \Phi$ for all $\alpha < \beta$ since for any $0 < \epsilon < \beta - \alpha$, we have $q_{\beta-\epsilon} \in \Phi$ which implies that there exists a prediction $a \in \mathcal{A}$ such that $\ell_a(y) \leq -\frac{1}{\eta}\ln q_{\beta-\epsilon}(y) \leq -\frac{1}{\eta}\ln q_\alpha(y)$ for all y. Hence $-\frac{1}{\eta}\ln q_\alpha \in \mathcal{S}$, and $q_\alpha \in \Phi$. But now

$$\lim_{\alpha\uparrow\beta}\|q_\beta - q_\alpha\| = \lim_{\alpha\uparrow\beta}(\beta-\alpha)\|q\| \leq \lim_{\alpha\uparrow\beta}(\beta-\alpha) = 0,$$

so the assumption that Φ is closed implies that $q_\beta \in \Phi$, and hence $q_\beta \neq q$, showing that $\beta < 1$.

Finally, to prove that q_β lies on the boundary of Φ, consider a ball $B_\epsilon = \{r \in \Phi \mid \|r - q_\beta\| < \epsilon\}$ of arbitrary radius $\epsilon \in (0, 1-\beta]$. This ball contains the point $q_{\beta+\epsilon/2}$, which lies outside of Φ by definition of β. Hence B_ϵ is not contained in Φ for any ϵ, and consequently q_β must lie on the boundary of Φ. □

A.5 Proofs of Theorem 8 and Corollary 9

For $\eta > 0$, define

$$h_\eta(f, f^*) = \frac{1}{\eta}\Big(1 - \mathbf{E}\left[\frac{e^{-\eta\ell(Y,f(X))}}{e^{-\eta\ell(Y,f^*(X))}}\right]\Big).$$

The letter h comes from the special case of log-loss, $\mathcal{X} = \{x\}$ a singleton, and a correct model $\mathcal{F}$ that includes the true distribution $P^*(Y|X = x)$, because in this case $h_{1/2}$ is the squared Hellinger distance.

Also define the positive, continuous, increasing function $\phi(a) = (e^a - a - 1)/a^2$ for $a \neq 0$ and $\phi(0) = 1/2$.

We need the following lemma, which is similar to Lemma 8.2 by Audibert [27] and to item (4) of Proposition 1.2 by Zhang [21].

Lemma 10. *Suppose* $|\ell(Y, f(X)) - \ell(Y, f^*(X))| \leq V$ *(a.s.) for* $V < \infty$. *Then for any* $\eta > 0$ *there exists* $c_{\eta,f} \in [\phi(-\eta V), \phi(\eta V)]$ *such that*

$$d(f, f^*) = h_\eta(f, f^*) + c_{\eta,f}\,\eta V(f, f^*).$$

Proof. Let $Z = \ell(Y, f(X)) - \ell(Y, f^*(X)) \in [-V, V]$. We need to show

$$\mathbf{E}[Z] = \frac{1}{\eta}(1 - \mathbf{E}[e^{-\eta Z}]) + c_{\eta,f}\,\eta\,\mathbf{E}[Z^2]. \tag{13}$$

Suppose $\mathbf{E}[Z^2] = 0$. Then $Z = 0$ (a.s.), and (13) is satisfied for any constant $c_{\eta,f}$. Otherwise (13) may be rewritten as

$$\mathbf{E}\left[\frac{(\eta Z)^2}{\mathbf{E}[(\eta Z)^2]}\cdot\phi(-\eta Z)\right] = c_{\eta,f}.$$

Recognising the left-hand side as the expectation of $\phi(-\eta Z)$ under the distribution with density $(\eta Z)^2 \mathrm{d}P^* / \mathbf{E}[(\eta Z)^2]$, its value must lie in the interval $[\min_z \phi(-\eta z), \max_z \phi(-\eta z)]$. As ϕ is increasing, these extreme values are achieved at $z = -V$ and $z = V$, from which the lemma follows. □

Proof of Theorem 8. Although h_η is nonnegative when it equals the squared Hellinger distance, this property does not hold in general. In fact, we observe that η-stochastic mixability up to ϵ is equivalent to

$$h_\eta(f, f^*) \geq 0 \quad \text{for all } f \in \mathcal{F} \text{ such that } d(f, f^*) \geq \epsilon. \tag{14}$$

(Only if) Suppose the margin condition (7) holds with constants $\kappa \geq 1$ and $c_0 > 0$. Then Lemma 10 implies that

$$d(f, f^*) - h_\eta(f, f^*) \leq \phi(\eta V)\eta V(f, f^*) \leq \phi(\eta V)\eta c_0^{-1/\kappa} d(f, f^*)^{1/\kappa}. \tag{15}$$

Now let $\epsilon > 0$ be arbitrary. As the loss is bounded by V, we have $d(f, f^*) \leq V$. Hence for $\epsilon > V$ (14) is trivially satisfied. So assume without loss of generality that $\epsilon \leq V$, and let $\eta = C\epsilon^{\frac{\kappa-1}{\kappa}}$ for some constant $C \in (0, V^{-\frac{\kappa-1}{\kappa}}]$ to be determined later. Then $\eta \leq 1$, so that the fact that ϕ is increasing implies $\phi(\eta V) \leq \phi(V)$. Now for any $f \in \mathcal{F}$ such that $d(f, f^*) \geq \epsilon$ we have

$$\phi(\eta V)\eta c_0^{-1/\kappa} \leq \phi(V) c_0^{-1/\kappa} C \epsilon^{\frac{\kappa-1}{\kappa}} \leq \phi(V) c_0^{-1/\kappa} C d(f, f^*)^{\frac{\kappa-1}{\kappa}}.$$

Combining this with (15), we find

$$\begin{aligned} d(f, f^*) - h_\eta(f, f^*) &\leq \phi(V) c_0^{-1/\kappa} C d(f, f^*) \\ h_\eta(f, f^*) &\geq \big(1 - \phi(V) c_0^{-1/\kappa} C\big) d(f, f^*). \end{aligned}$$

Taking $C = \min\left\{\frac{c_0^{1/\kappa}}{\phi(V)}, \frac{1}{V^{(\kappa-1)/\kappa}}\right\}$ such that $1 - \phi(V)c_0^{-1/\kappa} C \geq 0$, and using $d(f, f^*) \geq 0$, we find that $h_\eta(f, f^*) \geq 0$ as required. This shows that the margin condition implies η-stochastic mixability up to ϵ for $\eta = C\epsilon^{(\kappa-1)/\kappa}$.

(If) Suppose the margin condition does not hold for κ. That is, for every $c_0 > 0$ there exists $f_{c_0} \in \mathcal{F}$ such that

$$c_0 V(f_{c_0}, f^*)^\kappa > d(f_{c_0}, f^*).$$

We will show that for every $C > 0$ there exists $\epsilon > 0$ such that (14) with $\eta = C\epsilon^{(\kappa-1)/\kappa}$ is violated. Let $C > 0$ be arbitrary and take $\epsilon = d(f_{c_0}, f^*) \leq V$ for some $c_0 > 0$ to be determined later. Then $\eta \leq CV^{(\kappa-1)/\kappa}$ so that $\phi(-\eta V) \geq \phi(-CV^{2-1/\kappa})$ and hence Lemma 10 implies that

$$\begin{aligned} d(f_{c_0}, f^*) - h_\eta(f_{c_0}, f^*) &\geq \phi(-\eta V)\eta V(f_{c_0}, f^*) > \phi(-\eta V)\eta c_0^{1/\kappa} d(f_{c_0}, f^*)^{1/\kappa} \\ \epsilon - h_\eta(f_{c_0}, f^*) &> \phi(-CV^{2-1/\kappa})\eta c_0^{1/\kappa}\epsilon^{1/\kappa} = \phi(-CV^{2-1/\kappa}) c_0^{1/\kappa} C\epsilon \\ h_\eta(f_{c_0}, f^*) &< \big(1 - \phi(-CV^{2-1/\kappa}) c_0^{1/\kappa} C\big)\epsilon. \end{aligned}$$

Choosing $c_0 \geq \big(\phi(-CV^{2-1/\kappa})C\big)^{-\kappa}$ gives $1 - \phi(-CV^{2-1/\kappa})c_0^{1/\kappa} C \leq 0$ and so we find that $h_\eta(f_{c_0}, f^*) < 0$ for $f_{c_0} \in \mathcal{F}$ such that $d(f_{c_0}, f^*) = \epsilon$. This violates (14), as was to be shown. □

Lemma 11. *Suppose the margin condition* (7) *is satisfied for some constants* $c_0 > 0$ *and* $1 \leq \kappa < \infty$*. Then the loss of* f^* *is almost surely unique. That is, if* $\mathbf{E}[\ell(Y, g^*(X))] = \mathbf{E}[\ell(Y, f^*(X))] = \min_{f \in \mathcal{F}} \mathbf{E}[\ell(Y, f(X))]$*, then* $\ell(Y, g^*(X)) = \ell(Y, f^*(X))$ *almost surely.*

Proof. We have $d(g^*, f^*) = 0$, and hence (7) implies that $V(g^*, f^*) = 0$, from which the lemma follows. □

Proof of Corollary 9. If $(\ell, \mathcal{F}, P^*)$ is stochastically mixable, then the margin condition (7) holds with $\kappa = 1$ by Theorem 8. Conversely, if (7) holds with $\kappa = 1$ then Theorem 8 implies that $(\ell, \bigcup_{\epsilon > 0} \mathcal{F}_\epsilon, P^*)$ is stochastically mixable, which by Lemma 11 implies stochastic mixability of $(\ell, \mathcal{F}, P^*)$. □