Supplement to: Loss factorization, weakly supervised learning and label noise robustness

A Proofs

A.1 Proof of Lemma 5

We need to show the double implication that defines sufficiency for y.
$\Rightarrow$) By Factorization Theorem (3), $R_{\mathcal{S},\ell}(h) - R_{\mathcal{S}',\ell}(h)$ is label independent only if the odd part cancels out.
$\Leftarrow$) If $\boldsymbol{\mu}_{\mathcal{S}} = \boldsymbol{\mu}'_{\mathcal{S}}$ then $R_{\mathcal{S},\ell}(h) - R_{\mathcal{S}',\ell}(h)$ is independent of the label, because the label only appears in the mean operator due to Factorization Theorem (3).

A.2 Proof of Lemma 6

Consider the class of LOLs satisfying $\ell(x) - \ell(-x) = 2ax$. For any element of the class, define $\ell_e(x) = \ell(x) - ax$, which is even. In fact we have

$$\ell_e(-x) = \ell(-x) + ax = \ell(x) - 2ax + ax = \ell(x) - ax = \ell_e(x) \ .$$

A.3 Proof of Theorem 7

We start by proving two helper Lemmas. The next one provides a bound to the Rademacher complexity computed on the sample $\mathcal{S}_{2x} \doteq \{(\boldsymbol{x}_i, \sigma), i \in [m], \forall \sigma \in \mathcal{Y}\}$.

Lemma 1 *Suppose m even. Suppose $\mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq X\}$ be the observations space, and $\mathcal{H} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq B\}$ be the space of linear hypotheses. Let $\mathcal{Y}^{2m} \doteq \times_{j\in[2m]}\mathcal{Y}$. Then the empirical Rademacher complexity*

$$\mathcal{R}(\mathcal{H} \circ \mathcal{S}_{2x}) \doteq \mathbb{E}_{\sigma\sim\mathcal{Y}^{2m}}\left[\sup_{\boldsymbol{\theta}\in\mathcal{H}} \frac{1}{2m}\sum_{i\in[2m]} \sigma_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\right]$$

of $\mathcal{H}$ on $\mathcal{S}_{2x}$ satisfies:

$$\mathcal{R}(\mathcal{H} \circ \mathcal{S}_{2x}) \quad\leq\quad v \cdot \frac{BX}{\sqrt{2m}} \ , \tag{1}$$

with $v \doteq \frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} - \frac{1}{m}}$.

Proof Suppose without loss of generality that $\boldsymbol{x}_i = \boldsymbol{x}_{m+i}$. The proof relies on the observation that $\forall \boldsymbol{\sigma} \in \mathcal{Y}^{2m}$,

$$\begin{aligned}
\arg\sup_{\boldsymbol{\theta}\in\mathcal{H}} \{\mathbb{E}_{\mathcal{S}}[\sigma(\boldsymbol{x})\langle\boldsymbol{\theta},\boldsymbol{x}\rangle]\} &= \frac{1}{2m}\arg\sup_{\boldsymbol{\theta}\in\mathcal{H}}\left\{\sum_i \sigma_i\langle\boldsymbol{\theta},\boldsymbol{x}_i\rangle\right\} \\
&= \frac{\sup_{\mathcal{H}}\|\boldsymbol{\theta}\|_2}{\|\sum_i \sigma_i\boldsymbol{x}_i\|_2}\sum_i \sigma_i\boldsymbol{x}_i \ . \qquad (2)
\end{aligned}$$

So,

$$\begin{aligned}
\mathcal{R}(\mathcal{H}\circ\mathcal{S}_{2x}) &= \mathbb{E}_{\mathcal{Y}^{2m}}\sup_{h\in\mathcal{H}}\{\mathbb{E}_{\mathcal{S}_{2x}}[\sigma(\boldsymbol{x})h(\boldsymbol{x})]\} \\
&= \frac{\sup_{\mathcal{H}}\|\theta\|_2}{2m}\cdot\mathbb{E}_{\mathcal{Y}^{2m}}\left[\frac{\left(\sum_{i=1}^{2m}\sigma_i\boldsymbol{x}_i\right)^\top\left(\sum_{i=1}^{2m}\sigma_i\boldsymbol{x}_i\right)}{\|\sum_{i=1}^{2m}\sigma_i\boldsymbol{x}_i\|_2}\right] \\
&= \sup_{\mathcal{H}}\|\theta\|_2\cdot\mathbb{E}_{\mathcal{Y}^{2m}}\left[\frac{1}{2m}\cdot\left\|\sum_{i=1}^{2m}\sigma_i\boldsymbol{x}_i\right\|_2\right] \ . \qquad (3)
\end{aligned}$$

Now, remark that whenever $\sigma_i = -\sigma_{m+i}$, $\boldsymbol{x}_i$ disappears in the sum, and therefore the max norm for the sum may decrease as well. This suggests to split the 2^{2m} assignations into 2^m groups of size 2^m, ranging over the possible number of observations taken into account in the sum. They can be factored by a weighted sum of contributions of each subset of indices $\mathcal{I} \subseteq [m]$ ranging over the non-duplicated observations:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}^{2m}}\left[\frac{1}{m}\cdot\left\|\sum_{i=1}^{2m}\sigma_i\boldsymbol{x}_i\right\|_2\right] &= \frac{1}{2^{2m}}\sum_{\mathcal{I}\subseteq[m]}\frac{2^{m-|\mathcal{I}|}}{2m}\cdot\sum_{\boldsymbol{\sigma}\in\mathcal{Y}^{|\mathcal{I}|}}\sqrt{2}\left\|\sum_{i\in\mathcal{I}}\sigma_i\boldsymbol{x}_i\right\|_2 \ . \qquad (4) \\
&= \frac{\sqrt{2}}{2^m}\sum_{\mathcal{I}\subseteq[m]}\frac{1}{2m}\cdot\underbrace{\frac{1}{2^{|\mathcal{I}|}}\cdot\sum_{\boldsymbol{\sigma}\in\mathcal{Y}^{|\mathcal{I}|}}\left\|\sum_{i\in\mathcal{I}}\sigma_i\boldsymbol{x}_i\right\|_2}_{u_{|\mathcal{I}|}} \ . \qquad (5)
\end{aligned}$$

The $\sqrt{2}$ factor appears because of the fact that we now consider only the observations of $\mathcal{S}$. Now, for any *fixed* $\mathcal{I}$, we renumber its observations in $[|\mathcal{I}|]$ for simplicity, and observe that, since $\sqrt{1+x} \leq 1 + x/2$,

$$\begin{aligned}
u_{|\mathcal{I}|} &= \frac{1}{2^{|\mathcal{I}|}} \sum_{\boldsymbol{\sigma}\in\mathcal{Y}^{|\mathcal{I}|}} \sqrt{\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2 + \sum_{i_1\neq i_2} \sigma_{i_1}\sigma_{i_2}\boldsymbol{x}_{i_1}^\top \boldsymbol{x}_{i_2}} && (6)\\
&= \frac{\sqrt{\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2}}{2^{|\mathcal{I}|}} \sum_{\boldsymbol{\sigma}\in\mathcal{Y}^{|\mathcal{I}|}} \sqrt{1 + \frac{\sum_{i_1\neq i_2} \sigma_{i_1}\sigma_{i_2}\boldsymbol{x}_{i_1}^\top \boldsymbol{x}_{i_2}}{\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2}} && (7)\\
&\leq \frac{\sqrt{\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2}}{2^{|\mathcal{I}|}} \sum_{\boldsymbol{\sigma}\in\mathcal{Y}^{|\mathcal{I}|}} \left(1 + \frac{\sum_{i_1\neq i_2} \sigma_{i_1}\sigma_{i_2}\boldsymbol{x}_{i_1}^\top \boldsymbol{x}_{i_2}}{2\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2}\right) && (8)\\
&= \sqrt{\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2} + \frac{1}{2^{|\mathcal{I}|}\cdot 2\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2} \cdot \sum_{\boldsymbol{\sigma}\in\mathcal{Y}^{|\mathcal{I}|}} \sum_{i_1\neq i_2} \sigma_{i_1}\sigma_{i_2}\boldsymbol{x}_{i_1}^\top \boldsymbol{x}_{i_2} && (9)\\
&= \sqrt{\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2} + \frac{1}{2^{|\mathcal{I}|}\cdot 2\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2} \cdot \sum_{i_1\neq i_2} \boldsymbol{x}_{i_1}^\top \boldsymbol{x}_{i_2} \cdot \underbrace{\left(\sum_{\boldsymbol{\sigma}\in\mathcal{Y}^{|\mathcal{I}|}} \sigma_{i_1}\sigma_{i_2}\right)}_{=0} && (10)\\
&= \sqrt{\sum_{i\in\mathcal{I}} \|\boldsymbol{x}_i\|_2^2} && (11)\\
&\leq \sqrt{|\mathcal{I}|}\cdot X\enspace . && (12)
\end{aligned}$$

Plugging this in eq. (5) yields

$$\frac{1}{X}\cdot \mathbb{E}_{\mathcal{Y}^{2m}}\left[\frac{1}{m}\cdot\left\|\sum_{i=1}^{2m}\sigma_i\boldsymbol{x}_i\right\|_2\right] \leq \frac{\sqrt{2}}{2^m}\sum_{k=0}^{m}\frac{\sqrt{k}}{2m}\binom{m}{k}\enspace . \tag{13}$$

Since m is even:

$$\mathbb{E}_{\mathcal{Y}^{2m}}\left[\frac{1}{2m}\cdot\left\|\sum_{i=1}^{2m}\sigma_i\boldsymbol{x}_i\right\|_2\right] \leq \frac{\sqrt{2}}{2^m}\sum_{k=0}^{(m/2)-1}\frac{\sqrt{k}}{2m}\binom{m}{k} + \frac{\sqrt{2}}{2^m}\sum_{k=m/2}^{m}\frac{\sqrt{k}}{2m}\binom{m}{k}\enspace . \tag{14}$$

Notice that the left one trivially satisfies

$$\begin{aligned}
\frac{\sqrt{2}}{2^m}\sum_{k=0}^{(m/2)-1}\frac{\sqrt{k}}{2m}\binom{m}{k} &\leq \frac{\sqrt{2}}{2^m}\sum_{k=0}^{(m/2)-1}\frac{1}{2m}\cdot\sqrt{\frac{m-2}{2}}\binom{m}{k}\\
&= \frac{1}{2}\cdot\sqrt{\frac{1}{m}-\frac{2}{m^2}}\cdot\frac{1}{2^m}\sum_{k=0}^{(m/2)-1}\binom{m}{k}\\
&\leq \frac{1}{4}\cdot\sqrt{\frac{1}{m}-\frac{2}{m^2}} && (15)
\end{aligned}$$

Also, the right one satisfies:

$$\begin{aligned}
\frac{\sqrt{2}}{2^m}\sum_{k=m/2}^{m}\frac{\sqrt{k}}{2m}\binom{m}{k} &\leq \frac{\sqrt{2}}{2^m}\sum_{k=m/2}^{m}\frac{\sqrt{m}}{2m}\binom{m}{k}\\
&= \frac{1}{\sqrt{2m}}\cdot\frac{1}{2^m}\sum_{k=m/2}^{m}\binom{m}{k}\\
&= \frac{1}{2}\cdot\frac{1}{\sqrt{2m}}\enspace . && (16)
\end{aligned}$$

We get

$$\frac{1}{X} \cdot \mathbb{E}_{\mathcal{Y}^{2m}}\left[\frac{1}{m} \cdot \left\|\sum_{i=1}^{2m} \sigma_i \boldsymbol{x}_i\right\|_2\right] \leq \frac{1}{4} \cdot \sqrt{\frac{1}{m} - \frac{2}{m^2}} + \frac{1}{2} \cdot \sqrt{\frac{1}{2m}} \tag{17}$$

$$= \frac{1}{\sqrt{2m}} \cdot \left(\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} - \frac{1}{m}}\right) . \tag{18}$$

And finally:

$$\mathcal{R}(\mathcal{H} \circ \mathcal{S}_{2x}) \leq v \cdot \frac{BX}{\sqrt{2m}} , \tag{19}$$

with

$$v \doteq \frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} - \frac{1}{m}} , \tag{20}$$

as claimed. ■

The second Lemma is a straightforward application of McDiarmid 's inequality [McDiarmid, 1998] to evaluate the convergence of the empirical mean operator to its population counterpart.

Lemma 2 *Suppose $\mathbb{R}^d \supseteq \mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq X < \infty\}$ be the observations space. Then for any $\delta > 0$ with probability at least $1 - \delta$*

$$\|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2 \leq X \cdot \sqrt{\frac{d}{m} \log\left(\frac{d}{\delta}\right)} .$$

Proof Let $\mathcal{S}$ and $\mathcal{S}'$ be two learning samples that differ for only one example $(\boldsymbol{x}_i, y_i) \neq (\boldsymbol{x}_{i'}, y_{i'})$. Let first consider the one-dimensional case. We refer to the k-dimensional component of $\boldsymbol{\mu}$ with $\boldsymbol{\mu}^k$. For any $\mathcal{S}, \mathcal{S}'$ and any $k \in [d]$ it holds

$$\begin{aligned} \left|\boldsymbol{\mu}_{\mathcal{S}}^k - \boldsymbol{\mu}_{\mathcal{S}'}^k\right| &= \frac{1}{m}\left|\boldsymbol{x}_i^k y_i - \boldsymbol{x}_{i'}^k y_{i'}\right| \\ &\leq \frac{X}{m}\left|y_i - y_{i'}\right| \\ &\leq \frac{2X}{m} . \end{aligned}$$

This satisfies the bounded difference condition of McDiarmid's inequality, which let us write for any $k \in [d]$ and any $\epsilon > 0$ that

$$\mathbb{P}\left(\left|\boldsymbol{\mu}_{\mathcal{D}}^k - \boldsymbol{\mu}_{\mathcal{S}}^k\right| \geq \epsilon\right) \leq \exp\left(-\frac{m\epsilon^2}{2X^2}\right)$$

and the multi-dimensional case, by union bound

$$\mathbb{P}\left(\exists k \in [d] : \left|\boldsymbol{\mu}_{\mathcal{D}}^k - \boldsymbol{\mu}_{\mathcal{S}}^k\right| \geq \epsilon\right) \leq d \exp\left(-\frac{m\epsilon^2}{2X^2}\right) .$$

Then by negation

$$\mathbb{P}\left(\forall k \in [d] : \left|\boldsymbol{\mu}_{\mathcal{D}}^k - \boldsymbol{\mu}_{\mathcal{S}}^k\right| \leq \epsilon\right) \geq 1 - d \exp\left(-\frac{m\epsilon^2}{2X^2}\right) ,$$

which implies that for any $\delta > 0$ with probability $1-\delta$

$$X\sqrt{\frac{2}{m}\log\left(\frac{d}{\delta}\right)} \geq \|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\|_\infty \geq d^{-1/2}\,\|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2 \ .$$

This concludes the proof. ■

We now restate and prove Theorem 7.

Theorem 7 *Assume ℓ is a-LOL and L-Lipschitz. Suppose $\mathbb{R}^d \supseteq \mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq X < \infty\}$ be the observations space, and $\mathcal{H} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq B < \infty\}$ be the space of linear hypotheses. Let $c(X, B) \doteq \max_{y\in\mathcal{Y}} \ell(yXB)$. Let $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}\in\mathcal{H}} R_{\mathcal{S},\ell}(\boldsymbol{\theta})$. Then for any $\delta > 0$, with probability at least $1-\delta$*

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) \leq \left(\frac{\sqrt{2}+1}{4}\right)\cdot\frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2}\cdot\sqrt{\frac{1}{m}\log\left(\frac{1}{\delta}\right)} + 2|a|B\cdot\|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2 \ ,$$

or more explicitly

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) \leq \left(\frac{\sqrt{2}+1}{4}\right)\cdot\frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2}\sqrt{\frac{1}{m}\log\left(\frac{2}{\delta}\right)} + 2|a|XB\sqrt{\frac{d}{m}\log\left(\frac{2d}{\delta}\right)} \ .$$

Proof Let $\boldsymbol{\theta}^\star = \operatorname{argmin}_{\boldsymbol{\theta}\in\mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta})$. We have

$$\begin{aligned} R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) &= \frac{1}{2}R_{\mathcal{D}_{2x},\ell}(\hat{\boldsymbol{\theta}}) + a\langle\hat{\boldsymbol{\theta}}, \boldsymbol{\mu}_{\mathcal{D}}\rangle - \frac{1}{2}R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}^\star) - a\langle\boldsymbol{\theta}^\star, \boldsymbol{\mu}_{\mathcal{D}}\rangle && (21)\\ &= \frac{1}{2}\left(R_{\mathcal{D}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}^\star)\right) + a\langle\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star, \boldsymbol{\mu}_{\mathcal{D}}\rangle \\ &= \frac{1}{2}\left(R_{\mathcal{S}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta}^\star)\right) + a\langle\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star, \boldsymbol{\mu}_{\mathcal{D}}\rangle \\ &\quad + \left.\frac{1}{2}\left(R_{\mathcal{D}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{S}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}^\star) + R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta}^\star)\right)\ \right\}A_1 \ . && (22) \end{aligned}$$

Step 21 is obtained by the equality $R_{\mathcal{D},\ell}(\boldsymbol{\theta}) = \frac{1}{2}R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}) + a\langle\boldsymbol{\theta}, \boldsymbol{\mu}_{\mathcal{D}}\rangle$ for any $\boldsymbol{\theta}$. Now, rename Line 22 as A_1. Applying the same equality with regard to $\mathcal{S}$, we have

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) \leq \underbrace{R_{\mathcal{S},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{S},\ell}(\boldsymbol{\theta}^\star)}_{A_2} + \underbrace{a\langle\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star, \boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\rangle}_{A_3} + A_1 \ .$$

Now, A_2 is never more than 0 because $\hat{\boldsymbol{\theta}}$ is the minimizer of $R_{\mathcal{S},\ell}(\boldsymbol{\theta})$. From the Cauchy-Schwarz inequality and bounded models it holds true that

$$A_3 \leq |a|\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_2\cdot\left\|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\right\|_2 \leq 2|a|B\left\|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\right\|_2 \ . \tag{23}$$

We could treat A_1 by calling standard bounds based on Rademacher complexity on a sample with size $2m$ [Bartlett and Mendelson, 2002]. Indeed, since the complexity does not depend on labels, its value would be the same –modulo the change of sample size– for both $\mathcal{S}$ and $\mathcal{S}_{2x}$, as they are computed with same loss and observations. However, the special structure of $\mathcal{S}_{2x}$ allows us to obtain a tighter structural complexity term, due to some cancellation effect. The fact is proven by Lemma 1. In order to exploit it, we first observe that

$$\begin{aligned} A_1 &\leq \frac{1}{2}\left(R_{\mathcal{D}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{S}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}^\star) + R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta}^\star)\right) \\ &\leq \sup_{\boldsymbol{\theta}\in\mathcal{H}} |R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}) - R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta})| \end{aligned}$$

which by standard arguments [Bartlett and Mendelson, 2002] and the application of Lemma 1 gives a bound with probability at least $1-\delta$, $\delta > 0$

$$
\begin{aligned}
A_1 &\leq 2L \cdot \mathcal{R}(\mathcal{H} \circ \mathcal{S}_{2x}) + c(X,B)L \cdot \sqrt{\frac{1}{4m}\log\left(\frac{1}{\delta}\right)} \\
&\leq L \cdot \frac{\sqrt{2}+1}{\sqrt{2}} \cdot \frac{BX}{\sqrt{2m}} + c(X,B)L \cdot \sqrt{\frac{1}{4m}\log\left(\frac{1}{\delta}\right)}
\end{aligned}
$$

where $c(X,B) \doteq \max_{y\in\mathcal{Y}} \ell(yXB)$ and because $\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} - \frac{1}{m}} < \left(\frac{\sqrt{2}+1}{\sqrt{2}}\right)$, $\forall m > 0$. We combine the results and get with probability at least $1-\delta$, $\delta > 0$ that

$$
R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) \leq \left(\frac{\sqrt{2}+1}{2}\right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2} \cdot \sqrt{\frac{1}{m}\log\left(\frac{1}{\delta}\right)} + 2|a|B \cdot \|\boldsymbol{\mu}_\mathcal{D} - \boldsymbol{\mu}_\mathcal{S}\|_2 \ . \tag{24}
$$

This proves the first part of the statement. For the second one, we apply Lemma 2 that provides the probabilistic bound for the norm discrepancy of the mean operators. Consider that both statements are true with probability at least $1-\delta/2$. We write

$$
\begin{aligned}
\mathbb{P}\Bigg(&\left\{ R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) \leq \left(\frac{\sqrt{2}+1}{2}\right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2} \cdot \sqrt{\frac{1}{m}\log\left(\frac{2}{\delta}\right)} + 2|a|B \cdot \|\boldsymbol{\mu}_\mathcal{D} - \boldsymbol{\mu}_\mathcal{S}\|_2 \right\} \\
&\bigwedge \left\{ \|\boldsymbol{\mu}_\mathcal{D} - \boldsymbol{\mu}_\mathcal{S}\|_2 \leq X \cdot \sqrt{\frac{d}{m}\log\left(\frac{2d}{\delta}\right)} \right\} \Bigg) \geq 1 - \delta/2 - \delta/2 = 1-\delta \ ,
\end{aligned}
$$

and therefore with probability $1-\delta$

$$
R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) \leq \left(\frac{\sqrt{2}+1}{2}\right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2} \cdot \sqrt{\frac{1}{m}\log\left(\frac{2}{\delta}\right)} + 2|a|XB \cdot \sqrt{\frac{d}{m}\log\left(\frac{2d}{\delta}\right)} \ .
$$

■

A.4 Unbiased estimator for the mean operator with asymmetric label noise

Natarajan et al. [2013, Lemma 1] provides an unbiased estimator for a loss $\ell(x)$ computed on x of the form:

$$
\hat{\ell}(y\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle) \doteq \frac{(1-p_{-y}) \cdot \ell(\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle) + p_y \cdot \ell(-\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle)}{1-p_- - p_+}
$$

We apply it for estimating the mean operator instead of, from another perspective, for estimating a linear (unhinged) loss as in van Rooyen et al. [2015]. We are allowed to do so by the very result of the Factorization Theorem, since the noise corruption has effect on the linear-odd term of the loss only. The estimator of the sufficient statistic of a single example $y\boldsymbol{x}$ is

$$
\begin{aligned}
\hat{\boldsymbol{z}} &\doteq \frac{1 - p_{-y} + p_y}{1 - p_- - p_+} y\boldsymbol{x} \\
&= \frac{1 - (p_- - p_+)y}{1 - p_- - p_+} y\boldsymbol{x} \\
&= \frac{y - (p_- - p_+)}{1 - p_- - p_+} \boldsymbol{x} \ ,
\end{aligned}
$$

and its average, *i.e.* the mean operator estimator, is

$$\hat{\boldsymbol{\mu}}_{\mathcal{S}} \doteq \mathbb{E}_{\mathcal{S}}\left[\frac{y-(p_- + p_+)}{1-p_- - p_+}\boldsymbol{x}\right] \ ,$$

such that in expectation over the noisy distribution it holds $\mathbb{E}_{\tilde{\mathcal{D}}}[\hat{\boldsymbol{z}}] = \boldsymbol{\mu}_{\mathcal{D}}$. Moreover, the corresponding risk enjoys the same unbiasedness property. In fact

$$\begin{aligned}
\hat{R}_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}) &= \frac{1}{2}R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}) + \mathbb{E}_{\tilde{\mathcal{D}}}\left[a\langle\boldsymbol{\theta},\hat{\boldsymbol{z}}\rangle\right] \\
&= \frac{1}{2}R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}) + a\langle\boldsymbol{\theta},\hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}}\rangle \\
&= \frac{1}{2}R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}) + a\langle\boldsymbol{\theta},\boldsymbol{\mu}_{\mathcal{D}}\rangle \\
&= R_{\mathcal{D},\ell}(\boldsymbol{\theta}) \ ,
\end{aligned} \tag{25}$$

where we have also used the independency on labels (and therefore of label noise) of $R_{\mathcal{D}_{2x},\ell}$.

A.5 Proof of Theorem 8

This Theorem is a version of Theorem 7 applied to the case of asymmetric label noise. Those results differ in three elements. First, we consider the generalization property of a minimizer $\hat{\boldsymbol{\theta}}$ that is learnt on the corrupted sample $\tilde{\mathcal{S}}$. Second, the minimizer is computed on the basis of the unbiased estimator of $\hat{\boldsymbol{\mu}}_{\tilde{\mathcal{S}}}$ and not barely $\boldsymbol{\mu}_{\tilde{\mathcal{S}}}$. Third, as a consequence, Lemma 2 is not valid in this scenario. Therefore, we first prove a version of the bound for the mean operator norm discrepancy while considering label noise.

Lemma 3 *Suppose $\mathbb{R}^d \supseteq \mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq X < \infty\}$ be the observations space. Let $\tilde{\mathcal{S}}$ is a learning sample affected by asymmetric label noise with noise rates $(p_+, p_-) \in [0, 1/2)$. Then for any $\delta > 0$ with probability at least $1-\delta$*

$$\left\|\hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}} - \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{S}}}\right\|_2 \leq \frac{X}{1-p_- - p_+}\cdot\sqrt{\frac{d}{m}\log\left(\frac{d}{\delta}\right)} \ .$$

Proof Let $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{S}}'$ be two learning samples from the corrupted distribution $\tilde{\mathcal{D}}$ that differ for only one example $(\boldsymbol{x}_i, \tilde{y}_i) \neq (\boldsymbol{x}_{i'}, \tilde{y}_{i'})$. Let first consider the one-dimensional case. We refer to the k-dimensional component of $\boldsymbol{\mu}$ with $\boldsymbol{\mu}^k$. For any $\tilde{\mathcal{S}}, \tilde{\mathcal{S}}'$ and any $k \in [d]$ it holds

$$\begin{aligned}
\left|\hat{\boldsymbol{\mu}}^k_{\tilde{\mathcal{S}}} - \hat{\boldsymbol{\mu}}^k_{\tilde{\mathcal{S}}'}\right| &= \frac{1}{m}\left|\left(\frac{\tilde{y}_i - (p_- - p_+)}{1-p_- - p_+}\right)\boldsymbol{x}^k_i - \left(\frac{\tilde{y}_{i'} - (p_- - p_+)}{1-p_- - p_+}\right)\boldsymbol{x}^k_{i'}\right| \\
&= \frac{1}{m}\left|\frac{\tilde{y}_i\boldsymbol{x}^k_i}{1-p_- - p_+} - \frac{\tilde{y}_{i'}\boldsymbol{x}^k_{i'}}{1-p_- - p_+}\right| \\
&\leq \frac{X}{m(1-p_- - p_+)}\left|\tilde{y}_i - \tilde{y}_{i'}\right| \\
&\leq \frac{2X}{m(1-p_- - p_+)} \ .
\end{aligned}$$

This satisfies the bounded difference condition of McDiarmid's inequality, which let us write for any $k \in [d]$ and any $\epsilon > 0$ that

$$\mathbb{P}\left(\left|\hat{\boldsymbol{\mu}}^k_{\mathcal{D}} - \hat{\boldsymbol{\mu}}^k_{\mathcal{S}}\right| \geq \epsilon\right) \leq \exp\left(-(1-p_- - p_+)^2\frac{m\epsilon^2}{2X^2}\right)$$

and the multi-dimensional case, by union bound

$$\mathbb{P}\left(\exists k \in [d] : \left|\hat{\boldsymbol{\mu}}^k_{\mathcal{D}} - \hat{\boldsymbol{\mu}}^k_{\mathcal{S}}\right| \geq \epsilon\right) \leq d \exp\left(-(1-p_- - p_+)^2 \frac{m\epsilon^2}{2X^2}\right) \ .$$

Then by negation

$$\mathbb{P}\left(\forall k \in [d] : \left|\hat{\boldsymbol{\mu}}^k_{\mathcal{D}} - \hat{\boldsymbol{\mu}}^k_{\mathcal{S}}\right| \leq \epsilon\right) \geq 1 - d \exp\left(-(1-p_- - p_+)^2 \frac{m\epsilon^2}{2X^2}\right) \ ,$$

which implies that for any $\delta > 0$ with probability $1-\delta$

$$\frac{X}{(1-p_- - p_+)} \sqrt{\frac{2}{m} \log\left(\frac{d}{\delta}\right)} \geq \|\hat{\boldsymbol{\mu}}_{\mathcal{D}} - \hat{\boldsymbol{\mu}}_{\mathcal{S}}\|_\infty \geq d^{-1/2} \|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2 \ .$$

This concludes the proof. ■

The proof of Theorem 8 follows the structure of Theorem 7's and elements of Natarajan et al. [2013, Theorem 3]'s. Let $\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta} \in \mathcal{H}} \hat{R}_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^\star = \mathrm{argmin}_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta})$. We have

$$\begin{aligned}
R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) &= \hat{R}_{\tilde{\mathcal{D}},\ell}(\hat{\boldsymbol{\theta}}) - \hat{R}_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) && (26)\\
&= \frac{1}{2} R_{\mathcal{D}_{2x},\ell}(\hat{\boldsymbol{\theta}}) + a\langle \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}}\rangle - \frac{1}{2} R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}^\star) - a\langle \boldsymbol{\theta}^\star, \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}}\rangle \\
&= \frac{1}{2}\left(R_{\mathcal{D}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}^\star)\right) + a\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star, \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}}\rangle \\
&= \frac{1}{2}\left(R_{\mathcal{S}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta}^\star)\right) + a\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star, \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}}\rangle \\
&\quad + \left.\frac{1}{2}\left(R_{\mathcal{D}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{S}_{2x},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D}_{2x},\ell}(\boldsymbol{\theta}^\star) + R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta}^\star)\right)\ \right\} A_1 \ . && (27)
\end{aligned}$$

Step 26 is due to unbiasedness shown in Section A.4. Again, rename Line 27 as A_1, which this time is bounded directly by Theorem 7. Next, we proceed as within the proof of Theorem 7 but now exploiting the fact that $\frac{1}{2} R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta}) = \hat{R}_{\tilde{\mathcal{S}},\ell}(\boldsymbol{\theta}) - a\langle \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}}\rangle$

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) \leq \underbrace{\hat{R}_{\tilde{\mathcal{S}},\ell}(\hat{\boldsymbol{\theta}}) - \hat{R}_{\tilde{\mathcal{S}},\ell}(\boldsymbol{\theta}^\star)}_{A_2} + \underbrace{a\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star, \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}} - \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{S}}}\rangle}_{A_3} + A_1 \ .$$

Now, A_2 is never more than 0 because $\hat{\boldsymbol{\theta}}$ is the minimizer of $\hat{R}_{\tilde{\mathcal{S}},\ell}(\boldsymbol{\theta})$. From the Cauchy-Schwarz inequality and bounded models it holds true that

$$A_3 \leq |a| \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_2 \cdot \left\|\hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}} - \hat{\boldsymbol{\mu}}_{\mathcal{S}}\right\|_2 \leq 2|a|B \left\|\hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}} - \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{S}}}\right\|_2 \ , \tag{28}$$

for which we can call Lemma 3. Finally, by a union bound we get that for any $\delta > 0$ with probability $1-\delta$

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) \leq \left(\frac{\sqrt{2}+1}{2}\right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2} \sqrt{\frac{1}{m} \log\left(\frac{2}{\delta}\right)} + \frac{2|a|XB}{1-p_+ - p_-} \sqrt{\frac{d}{m} \log\left(\frac{2d}{\delta}\right)} \ .$$

A.6 Proof of Theorem 10

We now restate and prove Theorem 8. The reader might question the bound for the fact that the quantity on the right-hand side can change by rescaling $\boldsymbol{\mu}_{\mathcal{D}}$ by X, *i.e.* the max L_2 norm of observations in the space $\mathcal{X}$. Although, such transformation would affect ℓ-risks on the left-hand side as well, balancing the effect. With this in mind, we formulate the result without making explicit dependency on X.

Theorem 10 *Assume* $\{\boldsymbol{\theta} \in \mathcal{H} : ||\boldsymbol{\theta}||_2 \le B\}$. *Let* $(\boldsymbol{\theta}^\star, \tilde{\boldsymbol{\theta}}^\star)$ *respectively the minimizers of* $(R_{\mathcal{D},\ell}(\boldsymbol{\theta}), R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}))$ *in* $\mathcal{H}$. *Then every* a-LOL *is* ϵ-ALN. *That is*

$$R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star) \le 4|a|B\max(p_-, p_+) \cdot \|\boldsymbol{\mu}_{\mathcal{D}}\|_2 \ .$$

Moreover:
1. If $\|\boldsymbol{\mu}_{\mathcal{D}}\|_2 = 0$ *for* $\mathcal{D}$ *then every* LOL *is* ALN *for any* $\tilde{\mathcal{D}}$.
2. Suppose that ℓ *is also once differentiable and* γ*-strongly convex. Then* $\|\boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}^\star\|_2^2 \le 2\epsilon/\gamma$.

Proof The proof draws ideas from Manwani and Sastry [2013]. Let us first assume the noise to be symmetric, *i.e.* $p_+ = p_- = p$. For any $\boldsymbol{\theta}$ we have

$$\begin{aligned} R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}) &= (1-p)\left(R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) - R_{\mathcal{D},\ell}(\boldsymbol{\theta})\right) \\ &\quad + p\left(R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) - R_{\mathcal{D},\ell}(\boldsymbol{\theta}) + 2a\langle \boldsymbol{\theta}^\star - \boldsymbol{\theta}, \boldsymbol{\mu}_{\mathcal{D}}\rangle\right) && (29) \\ &\le \left(R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) - R_{\mathcal{D},\ell}(\boldsymbol{\theta})\right) + 4|a|Bp\|\boldsymbol{\mu}_{\mathcal{D}}\|_2 && (30) \\ &\le 4|a|Bp\|\boldsymbol{\mu}_{\mathcal{D}}\|_2 \ . && (31) \end{aligned}$$

We are working with LOLs, which are such that $\ell(x) = \ell(-x) + 2ax$ and therefore we can take Step 29. Step 30 follows from Cauchy-Schwartz inequality and bounded models. Step 31 is true because $\boldsymbol{\theta}^\star$ is the minimizer of $R_{\mathcal{D},\ell}(\boldsymbol{\theta})$. We have obtained a bound for any $\boldsymbol{\theta}$ and so for the supremum with regard to $\boldsymbol{\theta}$. Therefore:

$$\sup_{\boldsymbol{\theta}\in\mathcal{H}} \Big(R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta})\Big) = R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}) \ .$$

To lift the discussion to *asymmetric* label noise, risks have to be split into losses for negative and positive examples. Let $R_{\mathcal{D}^+,\ell}$ be the risk computed over the distribution of the positive examples $\mathcal{D}^+$ and $R_{\mathcal{D}^-,\ell}$ the one of the negatives, and denote the mean operators $\boldsymbol{\mu}_{\mathcal{D}^+}, \boldsymbol{\mu}_{\mathcal{D}^-}$ accordingly. Also, define the probability of positive and negative labels in $\mathcal{D}$ as $\pi_\pm = \mathbb{P}(y = \pm 1)$. The same manipulations for the symmetric case let us write

$$\begin{aligned} R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}) &= \pi_-\left(R_{\mathcal{D}^-,\ell}(\boldsymbol{\theta}^\star) - R_{\mathcal{D}^-,\ell}(\boldsymbol{\theta})\right) + \pi_+\left(R_{\mathcal{D}^+,\ell}(\boldsymbol{\theta}^\star) - R_{\mathcal{D}^+,\ell}(\boldsymbol{\theta})\right) \\ &\quad + 2ap_-\pi_-\langle \boldsymbol{\theta}^\star - \boldsymbol{\theta}, \boldsymbol{\mu}_{\mathcal{D}^-}\rangle + 2ap_+\pi_+\langle \boldsymbol{\theta}^\star - \boldsymbol{\theta}, \boldsymbol{\mu}_{\mathcal{D}^+}\rangle \\ &\le \left(R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) - R_{\mathcal{D},\ell}(\boldsymbol{\theta})\right) + 2a\langle \boldsymbol{\theta}^\star - \boldsymbol{\theta}, p_-\boldsymbol{\mu}_{\mathcal{D}^-} + p_+\boldsymbol{\mu}_{\mathcal{D}^+}\rangle \\ &\le 4|a|B \cdot \|p_-\pi_-\boldsymbol{\mu}_{\mathcal{D}^-} + p_+\pi_+\boldsymbol{\mu}_{\mathcal{D}^+}\|_2 \\ &\le 4|a|B\max(p_-, p_+) \cdot \|\pi_-\boldsymbol{\mu}_{\mathcal{D}^-} + \pi_+\boldsymbol{\mu}_{\mathcal{D}^+}\|_2 \\ &= 4|a|B\max(p_-, p_+) \cdot \|\boldsymbol{\mu}_{\mathcal{D}}\|_2 \ . \end{aligned}$$

Then, we conclude the proof by the same argument for the symmetric case. The first corollary is immediate. For the second, we first recall the definition of a function f strongly convex.

Definition 4 *A differentiable function* $f(x)$ *is* γ*-strongly convex if for all* $x, x' \in Dom(f)$ *we have*

$$f(x) - f(x') \ge \langle \nabla f(x'), x - x'\rangle + \frac{\gamma}{2}\|x - x'\|_2^2 \ .$$

If ℓ is differentiable once and γ-strongly convex in the $\boldsymbol{\theta}$ argument, so it the risk $R_{\tilde{D},\ell}$ by composition with linear functions. Notice also that $\nabla R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star) = 0$ because $\tilde{\boldsymbol{\theta}}^\star$ is the minimizer. Therefore:

$$\begin{aligned}\epsilon &\geq R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star) \\ &\geq \left\langle \nabla R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star), \boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}^\star \right\rangle + \frac{\gamma}{2}\left\|\boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}^\star\right\|_2^2 \\ &\geq \frac{\gamma}{2}\left\|\boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}^\star\right\|_2^2 \ ,\end{aligned}$$

which means that

$$\left\|\boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}^\star\right\|_2^2 \leq \frac{2\epsilon}{\gamma} \ .$$

■

A.7 Proof of Lemma 11

$$\begin{aligned}\mathbb{C}\mathrm{ov}_{\mathcal{S}}[\boldsymbol{x}, y] &= \mathbb{E}_{\mathcal{S}}[y\boldsymbol{x}] - \mathbb{E}_{\mathcal{S}}[y]\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}] \\ &= \boldsymbol{\mu}_{\mathcal{S}} - \left(\frac{1}{m}\sum_{i:y_i>0} 1 - \frac{1}{m}\sum_{i:y_i<0} 1\right)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}] \\ &= \boldsymbol{\mu}_{\mathcal{S}} - (2\pi_+ - 1)\,\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}] \ .\end{aligned}$$

The second statement follows immediately.

B Factorization of non linear-odd losses

When ℓ_o is not linear, we can find upperbounds in the form of affine functions. It suffices to be continuous and have asymptotes at $\pm\infty$.

Lemma 5 *Let the loss ℓ be continuous. Suppose that it has asymptotes at $\pm\infty$,* i.e. *there exist $c_1, c_2 \in \mathbb{R}$ and $d_1, d_2 \in \mathbb{R}$ such that*

$$\lim_{x\to+\infty} \ell(x) - c_1 x - d_1 = 0, \quad \lim_{x\to-\infty} \ell(x) - c_2 x - d_2 = 0$$

then there exists $q \in \mathbb{R}$ such that $\ell_o(x) \leq \frac{c_1+c_2}{2}x + q$.

Proof One can compute the limits at infinity of ℓ_o to get

$$\lim_{x\to+\infty} \ell_o(x) - \frac{c_1+c_2}{2}x = \frac{d_1-d_2}{2}$$

and

$$\lim_{x\to-\infty} \ell_o(x) - \frac{c_1+c_2}{2}x = \frac{d_2-d_1}{2} \ .$$

Then $q \doteq \sup\{\ell_o(x) - \frac{c_1+c_2}{2}x\} < +\infty$ as ℓ_o is continuous. Thus $\ell_o(x) - \frac{c_1+c_2}{2}x \leq q$. ■

The Lemma covers many cases of practical interest outside the class of LOLs, *e.g.* hinge, absolute and Huber losses. Exponential loss is the exception since $\ell_o(x) = -\sinh(x)$ cannot be bounded. Consider now hinge loss:

$\ell(x) = [1-x]_+$ is not differentiable in 1 nor proper [Reid and Williamson, 2010], however it is continuous with asymptotes at $\pm\infty$. Therefore, for any $\boldsymbol{\theta}$ its empirical risk is bounded as

$$R_{\mathcal{S},hinge}(\boldsymbol{\theta}) \leq \frac{1}{2} R_{\mathcal{S}_{2x},hinge}(\boldsymbol{\theta}) - \frac{1}{2}\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + q \ ,$$

since $c_1 = 0$ and $c_2 = 1$. An alternative proof of this result on hinge is provided next, giving the exact value of $q = 1/2$. The odd term for hinge loss is

$$\begin{aligned} \ell_o(x) &= \frac{1}{2}\left([1-x]_+ - [1+x]_+\right) \\ &= \frac{1}{4}\left(-2x + |1-x| - |1+x|\right) \end{aligned}$$

due to an arithmetic trick for the max function: $\max(a,b) = (a+b)/2 + |b-a|/2$. Then for any x

$$\begin{aligned} |1-x| &\leq |x| + 1, \\ |1+x| &\geq |x| - 1 \end{aligned}$$

and therefore

$$\ell_o(x) \leq \frac{1}{4}(-2x + |x| + 1 - |x| + 1) = \frac{1}{2}(1-x) \ .$$

We also provide a "if-and-only-if" version of Lemma 5 fully characterizing which family of losses can be upperbounded by a LOL.

Lemma 6 *Let $l : \mathbb{R} \to \mathbb{R}$ a continuous function. Then there exists $c_1, d_1, d_2 \in \mathbb{R}$ such that*

$$\limsup_{x\to+\infty} \ell_o(x) - c_1 x - d_1 = 0 \tag{32}$$

and

$$\limsup_{x\to-\infty} \ell_o(x) - c_1 x - d_2 = 0 \ , \tag{33}$$

if and only if there exists $q, q' \in \mathbb{R}$ such that $\ell_o(x) \leq q'x + q$ for every $x \in \mathbb{R}$.

Proof $\Rightarrow$) Suppose that such limits exist and they are zero for some c_1, d_1, d_2. Let prove that ℓ_o is bounded from above by a line.

$$q = \sup_{x\in\mathbb{R}} \{\ell_o(x) - c_1 x\} < \infty \ ,$$

because ℓ_o is continuous. So for every $x \in \mathbb{R}$

$$\ell_o(x) \leq c_1 x + q \ .$$

In particular we can take c_1 as the angular coefficient of the line.
$\Leftarrow$) Vice versa we proceed by contradiction. Suppose that there exists $q, q' \in \mathbb{R}$ such that ℓ_o is bounded from above by $\ell(x) = q'x + q$. Suppose in addition that the conditions on the asymptotes (32) and (33) are false. This implies either the existence of a sequence $x_n \to +\infty$ such that

$$\lim_{n\to\infty} \ell_o(x_n) - q'x_n \to \pm\infty \ ,$$

or the existence of another sequence $x'_n \to -\infty$

$$\lim_{n\to\infty} \ell_o(y_n) - q'x'_n \to \pm\infty \ .$$

On one hand, if at least one of these two limits is $+\infty$ then we already reach a contradiction, because $\ell_o(x)$ is supposed to be bounded from above by $\ell(x) = q'x + q$. Suppose on the other hand that $x_n \to +\infty$ is such that

$$\lim_{n\to+\infty} \ell_o(x_n) - q'x_n \to -\infty \ .$$

Then defining $x'_n = -x_n$ we have

$$\lim_{n\to+\infty} \ell_o(w_n) - mx'_n \to +\infty \ ,$$

and for the same reason as above we reach a contradiction. ■

C Factorization of square loss for regression

We have formulated the Factorization Theorem for classification problems. However, a similar property holds for regression with square loss: $f(\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle, y) = (\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle - y_i)^2$ factors as

$$\mathbb{E}_{\mathcal{S}}[(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle - y)^2] = \mathbb{E}_{\mathcal{S}}\left[\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle^2\right] + \mathbb{E}_{\mathcal{S}}\left[y^2\right] - 2\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \ .$$

Taking the minimizers on both sides we obtain

$$\begin{aligned} \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{S}}[f(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle, y)] &= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{S}}\left[\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle^2\right] - 2\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \|X^\top \boldsymbol{\theta}\|_2^2 - 2\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \ . \end{aligned}$$

D The role of LOLs in du Plessis et al. [2015]

Let $\pi_+ \doteq \mathbb{P}(y = 1)$ and let $\mathcal{D}_+$ and $\mathcal{D}_-$ respectively the set of positive and negative examples in $\mathcal{D}$. Consider first

$$\mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] = \pi_+ \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}_+}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] + (1-\pi_+)\mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}_-}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] \tag{34}$$

Then, it is also true that

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\ell(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] = \pi_+ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_+}\left[\ell(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] + (1-\pi_+)\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_-}\left[\ell(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] \ . \tag{35}$$

Now, solve Equation 34 for $(1-\pi_+)\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_-}\left[\ell(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] = (1-\pi_+)\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_-}\left[-\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right]$ and substitute it into Equation 35 so as to obtain:

$$\begin{aligned} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\ell(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] &= \pi_+ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_+}\left[\ell(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] + \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] - \pi_+ \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}_+}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] \\ &= \pi_+ \left(\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_+}\left[\ell(+\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] - \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}_+}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right]\right) + \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] \\ &= \frac{\pi_+}{2}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_+}\left[\ell_o(+\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] + \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] \ , \end{aligned} \tag{36}$$

by our usual definition of $\ell_o(x) = \frac{1}{2}(\ell(x) - \ell(-x))$. Recall that one of the goals of the authors is to conserve the convexity of this new crafted loss function. Then, du Plessis et al. [2015, Theorem 1] proceeds stating that when ℓ_o is convex, it must also be linear. And therefore they must focus on LOLs. The result of du Plessis et al. [2015, Theorem 1] is immediate from the point of view of our theory: in fact, an odd function can be convex or concave only if it also linear. The resulting expression based on the fact $\ell(x) - \ell(-x) = 2ax$ simplifies into

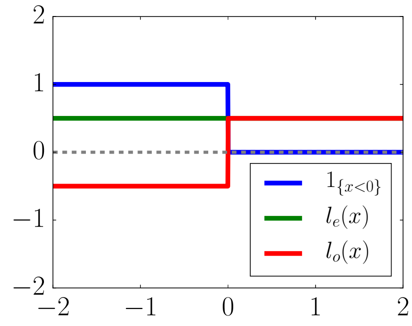
$$\begin{aligned} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\ell(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] &= a\pi_+ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_+}\left[y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle\right] + \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] \\ &= a\pi_+ \boldsymbol{\mu}_{\mathcal{D}_+} + \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mathcal{D}}\left[\ell(-\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)\right] \ . \end{aligned}$$

where $\boldsymbol{\mu}_{\mathcal{D}_+}$ is a mean operator computed on positive examples only. Notice how the second term is instead label independent, although it is not an even function as in the Factorization Theorem.

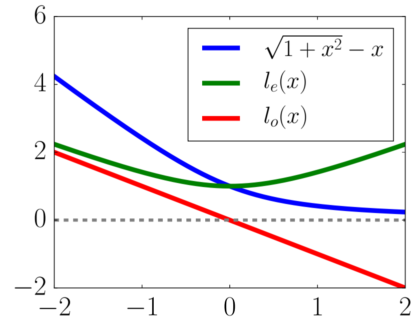
E Additional examples of loss factorization

	loss	even function ℓ_e	odd function l_o
generic	$\ell(x)$	$\frac{1}{2}(\ell(x)+\ell(-x))$	$\frac{1}{2}(\ell(x)-\ell(-x))$
01	$1\{x \leq 0\}$	$1-\frac{1}{2}\{x \neq 0\}$	$-\frac{1}{2}\operatorname{sign}(x)$
exponential	e^{-x}	$\cosh(x)$	$-\sinh(x)$
hinge	$[1-x]_+$	$\frac{1}{2}([1-x]_+ - [1-x]_+)$	$\frac{1}{2}([1-x]_+ - [1+x]_+)^\dagger$
LOL	$\ell(x)$	$\frac{1}{2}(\ell(x)+\ell(-x))$	$-ax$
ρ-loss	$\rho\lvert x\rvert - \rho x + 1$	$\rho\lvert x\rvert + 1$	$-\rho x \;\; (\rho \geq 0)$
unhinged	$1-x$	1	$-x$
perceptron	$\max(0,-x)$	$x\operatorname{sign}(x)$	$-x$
2-hinge	$\max(-x, 1/2\max(0,1-x))$	$\dagger\dagger$	$-x$
SPL	$a_l + l^\star(-x)/b_l$	$a_l + \frac{1}{2b_l}(l^\star(x)+l^\star(-x))$	$-x/(2b_l)$
logistic	$\log(1+e^{-x})$	$\frac{1}{2}\log(2+e^x+e^{-x})$	$-x/2$
square	$(1-x)^2$	$1+x^2$	$-2x$
Matsushita	$\sqrt{1+x^2}-x$	$\sqrt{1+x^2}$	$-x$

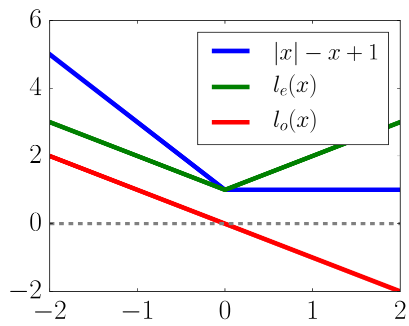
Table 1: Factorization of losses in light of Theorem 12. †The odd term of hinge loss is upperbounded by $(1-x)/2$ in B. $^{\dagger\dagger} = \max(-x, 1/2\max(0,1-x)) + \max(x, 1/2\max(0,1+x))$.



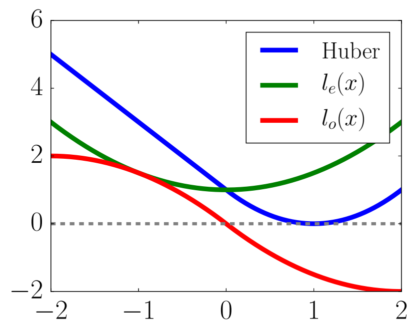
(a) 0-1 loss



(b) Matsushita loss



(c) ρ-loss, ρ=1



(d) Huber loss

References

C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 1–54. Springer Verlag, 1998.

P.-L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *NIPS*27*, 2013.

B. van Rooyen, A. K. Menon, and R. C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In *NIPS*29*, 2015.

N. Manwani and P. S. Sastry. Noise tolerance under risk minimization. *Cybernetics, IEEE Transactions on*, 43 (3):1146–1151, 2013.

M. D. Reid and R. C. Williamson. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

M C. du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *32 th ICML*, pages 1386–1394, 2015.