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), \( R_{S,\ell}(h) - R_{S',\ell}(h) \) is label independent only if the odd part cancels out.

\( \Leftarrow \) If \( \mu_S = \mu_S' \) then \( R_{S,\ell}(h) - R_{S',\ell}(h) \) 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 \( S_{2x} = \{(x_i, \sigma), i \in [m], \forall \sigma \in Y\} \).

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

\[ \mathcal{R}(\mathcal{H} \circ S_{2x}) \doteq \mathbb{E}_{\sigma \sim Y_{2m}} \left[ \sup_{\theta \in \mathcal{H}} \frac{1}{2m} \sum_{i \in [2m]} \sigma_i(\theta, x_i) \right] \]

of \( \mathcal{H} \) on \( S_{2x} \) satisfies:

\[ \mathcal{R}(\mathcal{H} \circ S_{2x}) \leq v \cdot \frac{BX}{\sqrt{2m}}, \]

(1)

with \( v = \frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2} - \frac{1}{m}}. \)
Proof Suppose without loss of generality that $x_i = x_{m+i}$. The proof relies on the observation that $\forall \sigma \in Y^{2m}$,

$$\arg \sup_{\theta \in \Theta} \{E_S[\sigma(x)\langle \theta, x \rangle]\} = \frac{1}{2m} \arg \sup_{\theta \in \Theta} \left\{\sum_i \sigma_i \langle \theta, x_i \rangle\right\} = \frac{\sup_{\theta \in \Theta} \|\theta\|_2}{\sum_i \sigma_i \|x_i\|_2} \sum_i \sigma_i x_i . \tag{2}$$

So,

$$\mathcal{R}(\mathcal{H} \circ S_2) = E_{Y^{2m}} \sup_{h \in \Theta} \{E_{S_2}[\sigma(x)h(x)]\}$$

$$= \sup_{\|\theta\|_2} \frac{1}{2m} \cdot E_{Y^{2m}} \left[\left(\sum_{i=1}^{2m} \sigma_i x_i\right)\top \left(\sum_{i=1}^{2m} \sigma_i x_i\right)\right] \frac{1}{\sum_{i=1}^{2m} \sigma_i ||x_i||_2}$$

$$= \sup_{\|\theta\|_2} \frac{1}{2m} \cdot \left[\frac{1}{2m} \sum_{i=1}^{2m} \sigma_i ||x_i||_2\right]. \tag{3}$$

Now, remark that whenever $\sigma_i = -\sigma_{m+i}$, $x_i$ disappears in the sum, and therefore the max norm for the sum may decrease as well. This suggests to split the $2^{2m}$ assignments 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 $J \subseteq [m]$ ranging over the non-duplicated observations:

$$E_{Y^{2m}} \left[\frac{1}{m} \cdot \left\|\sum_{i=1}^{2m} \sigma_i x_i\right\|_2\right] = \frac{1}{2^{2m}} \cdot \sum_{J \subseteq [m]} \left\{\frac{2^m}{2m} \cdot \sum_{\sigma \in Y^{|J|}} \sqrt{2} \left\|\sum_{i \in J} \sigma_i x_i\right\|_2\right\} . \tag{4}$$

$$= \sqrt{\frac{\sqrt{2}}{2^m}} \cdot \sum_{J \subseteq [m]} \left\{\frac{1}{2^{|J|}} \cdot \sum_{\sigma \in Y^{|J|}} \left\|\sum_{i \in J} \sigma_i x_i\right\|_2\right\} . \tag{5}$$
The $\sqrt{2}$ factor appears because of the fact that we now consider only the observations of $S$. Now, for any fixed $J$, we renumber its observations in $[|J|]$ for simplicity, and observe that, since $\sqrt{1 + x} \leq 1 + x/2$,

$$
\begin{align*}
\sqrt{\mathbb{E}}_{y^{2m}} \left[ 1/m \cdot \left\| \sum_{i=1}^{2m} \sigma_i x_i \right\|_2^2 \right] & \leq \frac{\sqrt{2}}{2m} \sum_{k=0}^{m} \sqrt{k} \left( \begin{array}{c} m \\frac{m}{2} \end{array} \right). \\

\end{align*}
$$

(6)

Since $m$ is even:

$$
\begin{align*}
\mathbb{E}_{y^{2m}} \left[ \frac{1}{2m} \cdot \left\| \sum_{i=1}^{2m} \sigma_i x_i \right\|_2^2 \right] & \leq \frac{\sqrt{2}}{2m} \sum_{k=0}^{(m/2)-1} \frac{\sqrt{k}}{2m} \left( \begin{array}{c} m \\frac{m}{2} \end{array} \right) + \frac{\sqrt{2}}{2m} \sum_{k=m/2}^{m} \frac{\sqrt{k}}{2m} \left( \begin{array}{c} m \\frac{m}{2} \end{array} \right).

\end{align*}
$$

(8)

Notice that the left one trivially satisfies

$$
\begin{align*}
\frac{\sqrt{2}}{2m} \sum_{k=0}^{(m/2)-1} \frac{1}{2m} \cdot \frac{m-2}{2m} \left( \begin{array}{c} m \\frac{m}{2} \end{array} \right) & \leq \frac{1}{2} \cdot \frac{1}{m} \cdot \frac{2}{m} \sum_{k=0}^{(m/2)-1} \left( \begin{array}{c} m \\frac{m}{2} \end{array} \right) \\
& \leq \frac{1}{4} \cdot \frac{1}{m} \cdot \frac{2}{m}.

\end{align*}
$$

(12)

Also, the right one satisfies:

$$
\begin{align*}
\frac{\sqrt{2}}{2m} \sum_{k=m/2}^{m} \frac{\sqrt{k}}{2m} \left( \begin{array}{c} m \\frac{m}{2} \end{array} \right) & \leq \frac{\sqrt{2}}{2m} \sum_{k=m/2}^{m} \frac{\sqrt{m}}{2m} \left( \begin{array}{c} m \\frac{m}{2} \end{array} \right) \\
& = \frac{1}{\sqrt{2m}} \cdot \frac{1}{2m} \sum_{k=m/2}^{m} \left( \begin{array}{c} m \\frac{m}{2} \end{array} \right) \\
& = \frac{1}{2} \cdot \frac{1}{\sqrt{2m}}.

\end{align*}
$$

(16)
We get
\[ \frac{1}{X} \mathbb{E}_{y^2} \left[ \frac{1}{m} \sum_{i=1}^{2m} \sigma_i x_i \right] \leq \frac{1}{4} \sqrt{\frac{1}{m} - \frac{2}{m^2}} + \frac{1}{2} \sqrt{\frac{1}{2m}} \]
(17)
\[ = \frac{1}{\sqrt{2m}} \left( \frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2} - \frac{1}{m}} \right) . \]
(18)
And finally:
\[ \mathcal{R}(\mathcal{H} \circ S_{2x}) \leq v \cdot \frac{BX}{\sqrt{2m}} , \]
(19)
with
\[ v = \frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2} - \frac{1}{m}} , \]
(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 \( \mathbb{R}^d \supseteq X = \{ x : \| x \|_2 \leq X < \infty \} \) be the observations space. Then for any \( \delta > 0 \) with probability at least \( 1 - \delta \)
\[ \| \mu_D - \mu_S \|_2 \leq X \cdot \frac{d \log \left( \frac{d}{\delta} \right)}{m} . \]
Proof Let \( S \) and \( S' \) be two learning samples that differ for only one example \((x_i, y_i) \neq (x_{i'}, y_{i'})\). Let first consider the one-dimensional case. We refer to the \( k \)-dimensional component of \( \mu \) with \( \mu^k \). For any \( S, S' \) and any \( k \in [d] \) it holds
\[ |\mu^k_S - \mu^k_{S'}| = \frac{1}{m} |x^k_i y_i - x^k_{i'} y_{i'}| \]
\[ \leq \frac{X}{m} |y_i - y_{i'}| \]
\[ \leq \frac{2X}{m} . \]
This satisfies the bounded difference condition of McDiarmid’s inequality, which let us write for any \( k \in [d] \) and any \( \epsilon > 0 \) that
\[ \mathbb{P} \left( |\mu^k_D - \mu^k_S| \geq \epsilon \right) \leq \exp \left( -\frac{m \epsilon^2}{2X^2} \right) \]
and the multi-dimensional case, by union bound
\[ \mathbb{P} \left( \exists k \in [d] : |\mu^k_D - \mu^k_S| \geq \epsilon \right) \leq d \exp \left( -\frac{m \epsilon^2}{2X^2} \right) . \]
Then by negation
\[ \mathbb{P} \left( \forall k \in [d] : |\mu^k_D - \mu^k_S| \leq \epsilon \right) \geq 1 - d \exp \left( -\frac{m \epsilon^2}{2X^2} \right) , \]
which implies that for any $\delta > 0$ with probability $1 - \delta$
\[
X \sqrt{\frac{2 \log \left( \frac{d}{\delta} \right)}{m}} \geq \|\mu_D - \mu_S\|_\infty \geq d^{-1/2} \|\mu_D - \mu_S\|_2 .
\]
This concludes the proof.

We now restate and prove Theorem 7.

**Theorem 7** Assume $\ell$ is a-LOL and L-Lipschitz. Suppose $\mathbb{R}^d \supseteq \mathcal{X} = \{x : \|x\|_2 \leq X < \infty\}$ be the observations space, and $\mathcal{H} = \{\theta : \|\theta\|_2 \leq B < \infty\}$ be the space of linear hypotheses. Let $c(X, B) = \max_{y \in Y} \ell(y X B)$. Let $\theta = \arg\min_{\theta \in \mathcal{H}} R_{S, \ell}(\theta)$. Then for any $\delta > 0$, with probability at least $1 - \delta$
\[
R_{D, \ell}(\hat{\theta}) - R_{D, \ell}(\theta^*) \leq \left( \frac{\sqrt{2} + 1}{4} \right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X, B)L}{2} \cdot \sqrt{\frac{1}{m} \log \left( \frac{1}{\delta} \right) + 2|a| B \cdot \|\mu_D - \mu_S\|_2} ,
\]
or more explicitly
\[
R_{D, \ell}(\hat{\theta}) - R_{D, \ell}(\theta^*) \leq \left( \frac{\sqrt{2} + 1}{4} \right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X, B)L}{2} \cdot \sqrt{\frac{1}{m} \log \left( \frac{2}{\delta} \right) + 2|a| XB \sqrt{\frac{d}{m} \log \left( \frac{2d}{\delta} \right)}}.
\]

**Proof** Let $\theta^* = \arg\min_{\theta \in \mathcal{H}} R_{D, \ell}(\theta)$. We have
\[
R_{D, \ell}(\hat{\theta}) - R_{D, \ell}(\theta^*) = \frac{1}{2} R_{D, \ell}(\hat{\theta}) + \theta^*, \mu_D) - \frac{1}{2} R_{D, \ell}(\theta^*, \mu_D)
\]
\[
= \frac{1}{2} \left( R_{D, \ell}(\hat{\theta}) - R_{D, \ell}(\theta^*) \right) + a(\hat{\theta} - \theta^*, \mu_D)
\]
\[
= \frac{1}{2} \left( R_{S, \ell}(\hat{\theta}) - R_{S, \ell}(\theta^*) \right) + a(\hat{\theta} - \theta^*, \mu_D)
\]
\[
+ \frac{1}{2} \left( R_{D, \ell}(\hat{\theta}) - R_{S, \ell}(\hat{\theta}) - R_{D, \ell}(\theta^*) + R_{S, \ell}(\theta^*) \right) A_1 .
\]
Step 21 is obtained by the equality $R_{D, \ell}(\hat{\theta}) = \frac{1}{2} R_{D, \ell}(\theta^*) + a(\theta^*, \mu_D)$ for any $\theta$. Now, rename Line 22 as
\[
A_2.
\]
Applying the same equality with regard to $S$, we have
\[
R_{D, \ell}(\hat{\theta}) - R_{D, \ell}(\theta^*) \leq \underbrace{R_{S, \ell}(\hat{\theta}) - R_{S, \ell}(\theta^*)}_{A_2} + \underbrace{a(\hat{\theta} - \theta^*, \mu_D - \mu_S)}_{A_2} + A_1 .
\]
Now, $A_2$ is never more than 0 because $\hat{\theta}$ is the minimizer of $R_{S, \ell}(\theta)$. From the Cauchy-Schwarz inequality and bounded models it holds true that
\[
A_3 \leq |a| \left\| \hat{\theta} - \theta^* \right\|_2 \cdot \left\| \mu_D - \mu_S \right\|_2 \leq 2|a| B \left\| \mu_D - \mu_S \right\|_2 .
\]
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 $S$ and $S_{2x}$, as they are computed with same loss and observations. However, the special structure of $S_{2x}$ 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
\[
A_1 \leq \frac{1}{2} \left( R_{D, \ell}(\hat{\theta}) - R_{S, \ell}(\theta) - R_{D, \ell}(\theta^*) + R_{S, \ell}(\theta^*) \right)
\]
\[
\leq \sup_{\theta \in \mathcal{H}} |R_{D, \ell}(\theta) - R_{S, \ell}(\theta)|
\]
which by standard arguments [Bartlett and Mendelson, 2002] and the application of Lemma 1 gives a bound with probability at least $1 - \delta$, $\delta > 0$

$$A_1 \leq 2L \cdot R(\mathcal{H} \circ S_{2x}) + c(X, B)L \cdot \sqrt{\frac{1}{4m} \log \left(\frac{1}{\delta}\right)}$$

$$\leq L \cdot \frac{\sqrt{2} + 1}{\sqrt{2}} \cdot BX \frac{X}{\sqrt{2m}} + c(X, B)L \cdot \sqrt{\frac{1}{4m} \log \left(\frac{1}{\delta}\right)}$$

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

$$R_{D,\ell}(\hat{\theta}) - R_{D,\ell}(\theta^*) \leq \left(\frac{\sqrt{2} + 1}{2}\right) \cdot \frac{XBL}{\sqrt{m}} + c(X, B)L \cdot \frac{1}{2} \sqrt{\frac{1}{m} \log \left(\frac{2}{\delta}\right)} + 2|a|B \cdot \|\mu_D - \mu_S\|_2 \cdot \sqrt{m} \log \left(\frac{2d}{\delta}\right). \quad (24)$$

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

$$\mathbb{P}\left\{ R_{D,\ell}(\hat{\theta}) - R_{D,\ell}(\theta^*) \leq \left(\frac{\sqrt{2} + 1}{2}\right) \cdot \frac{XBL}{\sqrt{m}} + c(X, B)L \cdot \frac{1}{2} \sqrt{\frac{1}{m} \log \left(\frac{2}{\delta}\right)} + 2|a|B \cdot \|\mu_D - \mu_S\|_2 \right\} \geq 1 - \delta / 2 - \delta / 2 = 1 - \delta,$$

and therefore with probability $1 - \delta$

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

\[\blacksquare\]

### 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:

$$\hat{\ell}(y(x, x_i)) = \frac{1 - p_y \cdot \ell(\theta, x_i) + p_y \cdot \ell(-\theta, x_i)}{1 - p_y - p_y}$$

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 $y$ is

$$\hat{z} = \frac{1 - p_y + p_y y x}{1 - p_y - p_y} = \frac{1 - (p_y - p_y) y}{1 - p_y - p_y} x \hat{y}$$
This satisfies the bounded difference condition of McDiarmid’s inequality, which let us write for any $\mu$:

$$\hat{\mu}_S = \mathbb{E}_S \left[ \frac{y - (p_- + p_+)}{1 - p_- - p_+} x \right],$$

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

$$\hat{R}_{D,\ell}^*(\theta) = \frac{1}{2} R_{D,\ell}^*(\theta) + \mathbb{E}_{\hat{\mu}}[a(\theta, \hat{z})]$$

and any $\ell > 0$

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{R}_{D,\ell}(\theta)$$

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

### 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 $\hat{\theta}$ that is learnt on the corrupted sample $\hat{S}$. Second, the minimizer is computed on the basis of the unbiased estimator of $\mu_D$ and not barely $\mu_{\hat{S}}$. 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 $\mathbb{R}^d \supset X = \{ x : \|x\|_2 \leq X < \infty \}$ be the observations space. Let $\hat{S}$ 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 $1 - \delta$

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

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

$$|\hat{\mu}^k_{\hat{S}} - \hat{\mu}^k_{\hat{S}'}| = \frac{1}{m} \left| \hat{y}_i \left( \frac{x_i - (p_- + p_+)}{1 - p_- - p_+} \right) - \hat{y}_i' \left( \frac{x_i' - (p_- + p_+)}{1 - p_- - p_+} \right) \right|$$

$$= \frac{1}{m} \left| \frac{\hat{y}_i x_i^k}{1 - p_- - p_+} - \frac{\hat{y}_i' x_i'^k}{1 - p_- - p_+} \right|$$

$$\leq \frac{X}{m(1 - p_- - p_+)} |\hat{y}_i - \hat{y}_i'|$$

$$\leq \frac{2X}{m(1 - p_- - p_+)}.$$

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

$$\mathbb{P} \left( |\hat{\mu}^k_D - \hat{\mu}^k_{\hat{S}}| \geq \epsilon \right) \leq \exp \left( -\frac{(1 - p_- - p_+)^2 m \epsilon^2}{2X^2} \right).$$
and the multi-dimensional case, by union bound
\[
\mathbb{P} \left( \exists k \in [d] : |\mathbf{\mu}_D^k - \mathbf{\mu}_S^k| \geq \epsilon \right) \leq d \exp \left( -\frac{(1 - p_- - p_+)^2 m \epsilon^2}{2X^2} \right).
\]
Then by negation
\[
\mathbb{P} \left( \forall k \in [d] : |\mathbf{\mu}_D^k - \mathbf{\mu}_S^k| \leq \epsilon \right) \geq 1 - d \exp \left( -\frac{(1 - p_- - p_+)^2 m \epsilon^2}{2X^2} \right),
\]
which implies that for any \( \delta > 0 \) with probability \( 1 - \delta \)
\[
\frac{X}{(1 - p_- - p_+)} \frac{2}{m \log (\frac{d}{\delta})} \geq \| \mathbf{\mu}_D - \mathbf{\mu}_S \|_\infty \geq d^{-1/2} \| \mathbf{\mu}_D - \mathbf{\mu}_S \|_2.
\]
This concludes the proof. \( \blacksquare \)

The proof of Theorem 8 follows the structure of Theorem 7’s and elements of Natarajan et al. [2013, Theorem 3]’s. Let \( \hat{\theta} = \arg\min_{\theta \in \mathcal{X}} R_{D,\ell}(\theta) \) and \( \theta^* = \arg\min_{\theta \in \mathcal{X}} R_{D,\ell}(\theta) \). We have
\[
R_{D,\ell}(\hat{\theta}) - R_{D,\ell}(\theta^*) = \hat{R}_{D,\ell}(\hat{\theta}) - \hat{R}_{D,\ell}(\theta^*) = \frac{1}{2} R_{D,\ell}(\hat{\theta}) + a\langle \hat{\theta}, \hat{\mu}_D \rangle - \frac{1}{2} R_{D,\ell}(\theta^*) - a\langle \theta^*, \hat{\mu}_D \rangle
\]
\[
= \frac{1}{2} \left( R_{D,\ell}(\hat{\theta}) - R_{D,\ell}(\theta^*) \right) + a\langle \hat{\theta} - \theta^*, \hat{\mu}_D \rangle
\]
\[
= \frac{1}{2} \left( R_{S,\ell}(\hat{\theta}) - R_{S,\ell}(\theta^*) \right) + a\langle \hat{\theta} - \theta^*, \hat{\mu}_D \rangle
\]
\[
+ \frac{1}{2} \left( R_{D,\ell}(\hat{\theta}) - R_{S,\ell}(\hat{\theta}) - R_{D,\ell}(\theta^*) + R_{S,\ell}(\theta^*) \right) \bigg) A_1. \quad (26)
\]
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 \( \frac{1}{2} R_{S,\ell}(\hat{\theta}) = \hat{R}_{S,\ell}(\hat{\theta}) - a\langle \hat{\theta}, \hat{\mu}_D \rangle \)
\[
R_{D,\ell}(\hat{\theta}) - R_{D,\ell}(\theta^*) \leq \hat{R}_{S,\ell}(\hat{\theta}) - \hat{R}_{S,\ell}(\theta^*) + a\langle \hat{\theta} - \theta^*, \hat{\mu}_D \rangle - \hat{R}_{S,\ell}(\hat{\theta}) + A_1. \quad (27)
\]
Now, \( A_2 \) is never more than 0 because \( \hat{\theta} \) is the minimizer of \( \hat{R}_{S,\ell}(\theta) \). From the Cauchy-Schwarz inequality and bounded models it holds true that
\[
A_3 \leq |a| \| \hat{\theta} - \theta^* \|_2 \cdot \| \hat{\mu}_D - \mu_S \|_2 \leq 2|a|B \| \hat{\mu}_D - \mu_S \|_2, \quad (28)
\]
for which we can call Lemma 3. Finally, by a union bound we get that for any \( \delta > 0 \) with probability \( 1 - \delta \)
\[
R_{D,\ell}(\hat{\theta}) - R_{D,\ell}(\theta^*) \leq \left( \frac{\sqrt{2} + 1}{2} \right) \frac{XBL}{\sqrt{m}} + \frac{c(X, B)L}{2} \frac{1}{m \log \left( \frac{2}{\delta} \right)} \sqrt{\frac{1}{m \log \left( \frac{2}{\delta} \right)} + \frac{2|a|X^2B}{1 - p_+ - p_-} \sqrt{\frac{d}{m \log \left( \frac{2d}{\delta} \right)}}}. \quad (29)
\]
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 \( \mathbf{\mu}_D \) by \( X \), i.e. the max \( L_2 \) norm of observations in the space \( \mathcal{X} \). Although, such transformation would affect \( \ell \)-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 $\{\theta \in \mathcal{H} : ||\theta||_2 \leq B\}$. Let $(\theta^*, \hat{\theta}^*)$ respectively the minimizers of $(R_{D,\ell}(\theta), R_{D,\ell}(\theta))$ in $\mathcal{H}$. Then every a-LOL is c-ALN. That is

$$R_{D,\ell}(\theta^*) - R_{D,\ell}(\hat{\theta}^*) \leq 4|a|B \max(p_-,p_+) \cdot \|\mu_D\|_2 .$$

Moreover:

1. If $\|\mu_D\|_2 = 0$ for $D$ then every LOL is ALN for any $\tilde{D}$.
2. Suppose that $\ell$ is also once differentiable and $\gamma$-strongly convex. Then $\|\theta^* - \hat{\theta}^*\|_2^2 \leq 2\epsilon/\gamma .$

**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 $\theta$ we have

$$R_{D,\ell}(\theta^*) - R_{D,\ell}(\theta) = (1 - p) (R_{D,\ell}(\theta^*) - R_{D,\ell}(\theta))$$

$$+ p (R_{D,\ell}(\theta^*) - R_{D,\ell}(\theta) + 2\langle \theta^* - \theta, \mu_D \rangle)$$

$$\leq \langle (R_{D,\ell}(\theta^*) - R_{D,\ell}(\theta)) + 4|a|Bp\|\mu_D\|_2$$

$$\leq 4|a|Bp\|\mu_D\|_2 .$$

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 $\theta^*$ is the minimizer of $R_{D,\ell}(\theta)$. We have obtained a bound for any $\theta$ and so for the supremum with regard to $\theta$. Therefore:

$$\sup_{\theta \in \mathcal{H}} \left( R_{D,\ell}(\theta^*) - R_{D,\ell}(\theta) \right) = R_{\tilde{D},\ell}(\theta^*) - R_{\tilde{D},\ell}(\hat{\theta}) .$$

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

$$R_{D,\ell}(\theta^*) - R_{D,\ell}(\theta) = \pi_-(R_{D^-,\ell}(\theta^*) - R_{D^-,\ell}(\theta)) + \pi_+(R_{D^+,\ell}(\theta^*) - R_{D^+,\ell}(\theta))$$

$$+ 2ap_\pi_-(\theta^* - \theta, \mu_{D^-}) + 2ap_\pi_+(\theta^* - \theta, \mu_{D^+})$$

$$\leq \langle (R_{D^+,\ell}(\theta^*) - R_{D^+,\ell}(\theta)) + 2\langle \theta^* - \theta, p_\pi_{D^-} + p_\pi_{D^+} \rangle$$

$$\leq 4|a|B \cdot \|p_\pi_{D^-} + p_\pi_{D^+} \cdot \|\mu_{D^-} + \mu_{D^+}\|_2$$

$$\leq 4|a|B \cdot \|\mu_{D^-} + \mu_{D^+}\|_2 .$$

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 $\gamma$-strongly convex if for all $x, x' \in \text{Dom}(f)$ we have

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

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

\[
\epsilon \geq R_{\tilde{D}, \ell}(\theta^*) - R_{\tilde{D}, \ell}(\tilde{\theta}^*) \\
\geq \langle \nabla R_{\tilde{D}, \ell}(\tilde{\theta}^*), \theta^* - \tilde{\theta}^* \rangle + \frac{\gamma}{2} \| \theta^* - \tilde{\theta}^* \|_2^2 \\
\geq \frac{\gamma}{2} \| \theta^* - \tilde{\theta}^* \|_2^2 ,
\]

which means that

\[
\| \theta^* - \tilde{\theta}^* \|_2^2 \leq \frac{2\epsilon}{\gamma}.
\]

**A.7 Proof of Lemma 11**

\[
\text{Cov}_S [x, y] = E_S [y|x] - E_S [y] E_S [x] \\
= \mu_S - \left( \frac{1}{m} \sum_{i: y_i > 0} 1 - \frac{1}{m} \sum_{i: y_i < 0} 1 \right) E_S [x] \\
= \mu_S - (2\pi_+ - 1) E_S [x].
\]

The second statement follows immediately.

**B Factorization of non linear-odd losses**

When \( \ell_o \) 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 \( \ell \) be continuous. Suppose that it has asymptotes at \( \pm \infty \), i.e. there exist \( c_1, c_2, d_1, d_2 \in \mathbb{R} \) 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 \( \ell_o \) 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 \( q = \sup \{ \ell_o(x) - \frac{c_1 + c_2}{2} x \} < +\infty \) as \( \ell_o \) 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 \( \theta \) its empirical risk is bounded as

\[
R_{S,\text{hinge}}(\theta) \leq \frac{1}{2} R_{S_{2\theta},\text{hinge}}(\theta) - \frac{1}{2} (\theta, \mu) + q ,
\]

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

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

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

\[
|1 - x| \leq |x| + 1, \\
|1 + x| \geq |x| - 1
\]

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 \( \ell_o \) is bounded from above by a line.

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

because \( \ell_o \) 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 \( \ell_o \) 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 \(x_n \rightarrow +\infty\) is such that

\[
\lim_{n \rightarrow +\infty} \ell_o(x_n) - q'x_n \rightarrow -\infty.
\]

Then defining \(x'_n = -x_n\) we have

\[
\lim_{n \rightarrow +\infty} \ell_o(w_n) - mx'_n \rightarrow +\infty,
\]

and for the same reason as above we reach a contradiction. \(\blacksquare\)

## 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: \(f(\langle \theta, x_\ell \rangle, y) = (\langle \theta, x_\ell \rangle - y_i)^2\) factors as

\[
E_S((\theta, x) - y)^2 = E_S[\langle \theta, x \rangle^2] + E_S[y^2] - 2\langle \theta, \mu \rangle.
\]

Taking the minimizers on both sides we obtain

\[
\arg\min_{\theta} E_S[f(\langle \theta, x \rangle, y)] = \arg\min_{\theta} E_S[\langle \theta, x \rangle^2] - 2\langle \theta, \mu \rangle = \arg\min_{\theta} \|X^T\theta\|_2^2 - 2\langle \theta, \mu \rangle.
\]

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

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

\[
E_{(x,\cdot) \sim \mathcal{D}}[\ell(-\langle \theta, x \rangle)] = \pi_+ E_{(x,\cdot) \sim \mathcal{D}_+}[\ell(-\langle \theta, x \rangle)] + (1 - \pi_+) E_{(x,\cdot) \sim \mathcal{D}_-}[\ell(-\langle \theta, x \rangle)]
\]

(34)

Then, it is also true that

\[
E_{(x,y) \sim \mathcal{D}}[\ell(y(\theta, x))] = \pi_+ E_{(x,y) \sim \mathcal{D}_+}[\ell(y(\langle \theta, x \rangle))] + (1 - \pi_+) E_{(x,y) \sim \mathcal{D}_-}[\ell(y(\langle \theta, x \rangle))] .
\]

(35)

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

\[
E_{(x,y) \sim \mathcal{D}}[\ell(y(\langle \theta, x \rangle))] = \pi_+ E_{(x,y) \sim \mathcal{D}_+}[\ell(y(\langle \theta, x \rangle))] + E_{(x,\cdot) \sim \mathcal{D}}[\ell(-\langle \theta, x \rangle)] - \pi_+ E_{(x,\cdot) \sim \mathcal{D}_+}[\ell(-\langle \theta, x \rangle)]
\]

\[
= \pi_+ \left(E_{(x,y) \sim \mathcal{D}_+}[\ell(+\langle \theta, x \rangle)] - E_{(x,\cdot) \sim \mathcal{D}_+}[\ell(-\langle \theta, x \rangle)]\right) + E_{(x,\cdot) \sim \mathcal{D}}[\ell(-\langle \theta, x \rangle)]
\]

\[
= \pi_+ \left(E_{(x,y) \sim \mathcal{D}_+}[\ell(\langle \theta, x \rangle)] + E_{(x,\cdot) \sim \mathcal{D}}[\ell(-\langle \theta, x \rangle)]\right) + \frac{\pi_+}{2} E_{(x,y) \sim \mathcal{D}_+}[\ell(\langle \theta, x \rangle)] .
\]

(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 \(\ell_o\) 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

\[
E_{(x,y) \sim \mathcal{D}}[\ell(y(\langle \theta, x \rangle))] = a\pi_+ E_{(x,y) \sim \mathcal{D}_+}[y(\langle \theta, x \rangle)] + E_{(x,\cdot) \sim \mathcal{D}}[\ell(-\langle \theta, x \rangle)]
\]

\[
= a\pi_+ \mu_{\mathcal{D}_+} + E_{(x,\cdot) \sim \mathcal{D}}[\ell(-\langle \theta, x \rangle)] .
\]

where \(\mu_{\mathcal{D}_+}\) 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

<table>
<thead>
<tr>
<th>loss</th>
<th>even function $\ell_e$</th>
<th>odd function $l_o$</th>
</tr>
</thead>
<tbody>
<tr>
<td>generic $\ell(x)$</td>
<td>$\frac{1}{2}(\ell(x) + \ell(-x))$</td>
<td>$\frac{1}{2}(\ell(x) - \ell(-x))$</td>
</tr>
<tr>
<td>$1{x \leq 0}$</td>
<td>$1 - \frac{1}{2}{x \neq 0}$</td>
<td>$-\frac{1}{2}\text{sign}(x)$</td>
</tr>
<tr>
<td>exponential $e^{-x}$</td>
<td>$\cosh(x)$</td>
<td>$-\sinh(x)$</td>
</tr>
<tr>
<td>$[1 - x]_+$</td>
<td>$\frac{1}{2}([1 - x]<em>+ - [1 - x]</em>+)$</td>
<td>$\frac{1}{2}([1 - x]<em>+ - [1 + x]</em>+)$†</td>
</tr>
<tr>
<td>LOL</td>
<td>$\ell(x)$</td>
<td>$\frac{1}{2}(\ell(x) + \ell(-x))$</td>
</tr>
<tr>
<td>$\rho$-loss</td>
<td>$\rho</td>
<td>x</td>
</tr>
<tr>
<td>unhinged &amp; $1 - x$ &amp; $1$ &amp; $-x$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hinge</td>
<td>$\max(0,-x)$</td>
<td>$x \text{sign}(x)$</td>
</tr>
<tr>
<td>2-hinge</td>
<td>$\max(-x, 1/2 \max(0, 1 - x))$</td>
<td>$\frac{1}{1+1}$††</td>
</tr>
<tr>
<td>SPL</td>
<td>$a_t + l_t(-x)/b_t$</td>
<td>$a_t + \frac{1}{2b_t}(l_t(x) + l_t(-x))$</td>
</tr>
<tr>
<td>logistic</td>
<td>$\log(1 + e^{-x})$</td>
<td>$\frac{1}{2} \log(2 + e^x + e^{-x})$</td>
</tr>
<tr>
<td>square</td>
<td>$(1 - x)^2$</td>
<td>$1 + x^2$</td>
</tr>
<tr>
<td>Matsushita</td>
<td>$\sqrt{1 + x^2} - x$</td>
<td>$\sqrt{1 + x^2}$</td>
</tr>
</tbody>
</table>

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

---

13
(a) 0-1 loss

(b) Matsushita loss

(c) $\rho$-loss, $\rho=1$

(d) Huber loss
References


