Goal: Guarantees for risk minimization, when the training labels (binary classification) are not reliable.

The label noise is:

1. random: labels are flipped iid.
2. class-conditioned: flip probability (noise rate) depends on the label (given by \( \rho_+ \) and \( \rho_- \)).

The class-conditional random noise model is given by:

\[
P(Y = + | Y = +) = \rho_+, \quad P(Y = + | Y = -) = \rho_-, \quad \text{and} \quad \rho_+ < 1, \rho_- < 1.
\]

Key observation: The Bayes optimal under the noisy distribution, \( \tilde{f} \) is given by

\[
\tilde{f}(x) = \operatorname{sign}(\tilde{\ell}(x) - 1/2) = \operatorname{sign}\left(f(x) - \frac{1}{2\rho_+ - \rho_-}\right).
\]

\( \tilde{f} \) and \( f \) differ only in the threshold.

Main Result: Method I

\( \tilde{f} \) obtained using (1) with the unbiased estimate \( \tilde{I} \) of loss \( I \).

\( \tilde{L} \): Lipschitz constant of \( f \).

**Main Result 1.** With probability at least \( 1 - \delta \),

\[
R(\tilde{\ell})_\alpha \leq \min_{R(\ell_\alpha)} \{4\ln(\mathcal{N}(F)) + 2\sqrt{\frac{\log(1/\delta)}{2n}}\}.
\]

Furthermore, if \( \tilde{f} \) is classification-calibrated, there exists a nondecreasing function \( \zeta \) with \( \zeta(0) = 0 \) such that

\[
R(\tilde{\ell}) - R(\ell) \leq \zeta\left(\min_{\alpha} \{R(\ell_\alpha) - \min_{\alpha} R(\ell_\alpha)\} + 4\ln(\mathcal{N}(F)) + 2\sqrt{\frac{\log(1/\delta)}{2n}}\right),
\]

where the RHS consists of approximation and estimation errors.

Main Results: Method II

Consider the risk of \( f \) w.r.t. the \( \alpha \)-weighted 0-1 loss under noisy distribution \( D_\alpha \):

\[
R(\tilde{\ell})_\alpha = E_{X,Y}(\tilde{\ell}_\alpha(X,Y) = 1 | \{X = 1 | Y = 1\}, \{X = 0 | Y = 0\}).
\]

**Main Result 2.** For the choices \( \alpha^* \) and \( A_0 \) above, there exists a nondecreasing function \( \zeta_0 \), with \( \zeta_0(0) = 0 \), such that, with probability at least \( 1 - \delta \):

\[
R(\tilde{\ell}_\alpha) - R(f) \leq A_\alpha \zeta_0\left(\min_{\alpha} \{R(f_\alpha) - \min_{\alpha} R(f_\alpha)\} + 4\ln(\mathcal{N}(F)) + 2\sqrt{\frac{\log(1/\delta)}{2n}}\right).
\]

\( \rho_+ - 1 = \rho_- \implies \alpha^* = 1/2 \implies \) the original loss suffices.

Method II: Weighted losses

For \( \alpha > 0 \), define \( \ell_\alpha \), as an \( \alpha \)-weighted margin loss function of the form:

\[
\ell_\alpha(x,y) = (1-\alpha)\ell(x,y) + \alpha \ell(1-x,y) = (1-\alpha)\ell(x,y) + \alpha \ell(1-x,y)\quad(\alpha > 0).
\]

Using \( \ell_\alpha \) in (1), we have the following ERM problem with noisy labels:

\[
\tilde{f}_\alpha = \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} \ell_\alpha(X_i, Y_i).
\]

For the hinge loss, letting \( \gamma \) be any bounded loss function. Then, if we define,

\[
\tilde{\ell}(x) = \log(\gamma(\ell(x))) = \log(\gamma(f(x) - 1/2)),
\]

\( \tilde{f}_\alpha \) is open if the biconjugate of \( \gamma \) is convex and twice differentiable.

Empirical Risk Minimization (ERM)

Let \( f: X \rightarrow \mathbb{R} \) be a real-valued decision function.

The risk of \( f \) w.r.t. the 0-1 loss is given by

\[
R(f) = E_X[1_{\{f(X) \neq Y\}}].
\]

Bayes optimal

\[
f^* = \arg\min_f R(f) - \log(2).\]

Empirical risk of the observed sample:

\[
\tilde{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \tilde{I}(X_i, Y_i)\]

\( \tilde{I} \) is the empirical risk under the “clean” distribution.

\[
\tilde{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \tilde{I}(X_i, Y_i).
\]

We want to bound \( R_\alpha(f) - \tilde{R}(f) \), i.e. the excess misclassification rate (under the clean distribution) of the empirical minimizer obtained using noisy data.

Method I: Unbiased Estimators

Let \( \ell(t, y) \) be any bounded loss function. Then, if we define,

\[
\tilde{I}(t, y) = \left(1 - \rho_+ - \rho_+\right) \ell(t, y) - \rho_+ \ell(t, y)
\]

we have, for any \( t, y \),

\[
E_T[\tilde{I}(t, Y)] = \tilde{I}(t, y).
\]

We can learn a predictor with strong risk bounds (Main Result 1) using the constructed \( \tilde{I} \) in ERM (1).

The idea is well-known in online learning, where examples arrive sequentially (Bounds akin to Main Result 1 can be obtained using a stochastic descent procedure).

Suppose \( \tilde{I}(t, y) \) is convex and twice differentiable almost everywhere in \( t \) (for every \( y \)) and also satisfies the symmetry property \( \tilde{I}(t, y) = \tilde{I}(t, -y) \), \( \forall t, \in T \), then \( \tilde{I} \) is also convex in \( t \).

Examples: the squared loss, the logistic loss, and the Huber loss.

Hinge loss does not satisfy the above property, but, with the class of bounded-norm hyperplanes \( \{w \in \mathbb{W} \} \), we can approximately minimize the non-convex \( F(W) = \tilde{R}(W) \) by minimizing the biconjugate \( F^* \).