Chebyshev-Cantelli PAC-Bayes-Bennett inequality for the weighted majority vote

Yi-Shan Wu, Andrés R. Masegosa, Stephan S. Lorenzen,
Christian Igel, Yevgeny Seldin

UNIVERSITY OF COPENHAGEN

Weighted Majority Vote

- Central technique for combining predictions of multiple classifiers (boosting, bagging, etc.)
- Wins most ML competitions.

Weighted Majority Vote

- Central technique for combining predictions of multiple classifiers (boosting, bagging, etc.)
- Wins most ML competitions.

Prediction rule

ρ-weighted majority vote MV_ρ predicts

$$\mathrm{MV}_\rho(X) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_\rho[\mathbb{1}(h(X) = y)].$$

Weighted Majority Vote

- Central technique for combining predictions of multiple classifiers (boosting, bagging, etc.)
- Wins most ML competitions.

Prediction rule

ρ-weighted majority vote MV_ρ predicts

$$\mathrm{MV}_\rho(X) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_\rho[\mathbb{1}(h(X) = y)].$$

Key Power

Cancellation of errors effect: Errors average out when

- errors of individual classifiers are independent
- individual classifiers have expected error less than 0.5

Standard Analysis

If a majority vote makes an error, at least a ρ-weighted half of the classifiers have made an error:

$$L(\mathrm{MV}_\rho) \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$$

Standard Analysis

If a majority vote makes an error, at least a ρ-weighted half of the classifiers have made an error:

$$L(\mathrm{MV}_\rho) \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$$

First Order Oracle Bound

By Markov's inequality $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{E}[Z]/\varepsilon$:

$$L(\mathrm{MV}_\rho) \leq 2\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] = 2\,\underbrace{\mathbb{E}_\rho[L(h)]}_{\text{Gibbs loss}}$$

Standard Analysis

If a majority vote makes an error, at least a ρ-weighted half of the classifiers have made an error:

$$L(\mathsf{MV}_\rho) \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$$

First Order Oracle Bound

By Markov's inequality $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{E}[Z]/\varepsilon$:

$$L(\mathsf{MV}_\rho) \leq 2\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] = 2\,\underbrace{\mathbb{E}_\rho[L(h)]}_{\text{Gibbs loss}}$$

Issues

Standard Analysis

If a majority vote makes an error, at least a ρ-weighted half of the classifiers have made an error:

$$L(\mathrm{MV}_\rho) \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$$

First Order Oracle Bound

By Markov's inequality $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{E}[Z]/\varepsilon$:

$$L(\mathrm{MV}_\rho) \leq 2\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] = 2\,\underbrace{\mathbb{E}_\rho[L(h)]}_{\text{Gibbs loss}}$$

Issues

- Ignores correlation of predictions (main power of MV)

Standard Analysis

If a majority vote makes an error, at least a ρ-weighted half of the classifiers have made an error:

$$L(\mathrm{MV}_\rho) \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$$

First Order Oracle Bound

By Markov's inequality $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{E}[Z]/\varepsilon$:

$$L(\mathrm{MV}_\rho) \leq 2\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] = 2\underbrace{\mathbb{E}_\rho[L(h)]}_{\text{Gibbs loss}}$$

Issues

- Ignores correlation of predictions (main power of MV)
- Optimization of corresponding PAC-Bayes bound degrades the test error [Lorenzen et al., 2019]

C-bounds [Lacasse et al., 2007, Germain et al., 2015, Laviolette et al., 2017]

$$\mathbb{P}(Z > 0.5) = \mathbb{P}(Z - \mathbb{E}[Z] \geq 0.5 - \mathbb{E}[Z])$$

by Chebyshev-Cantelli inequality, if $\mathbb{E}[Z] < 0.5$,

$$\begin{aligned} &\leq \frac{\mathbb{V}[Z]}{(0.5 - \mathbb{E}[Z])^2 + \mathbb{V}[Z]} \\ &= \frac{\mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2}{0.25 - \mathbb{E}[Z] + \mathbb{E}\left[Z^2\right]} \end{aligned}$$

since $V[Z] = \mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2$.

C-bounds [Lacasse et al., 2007, Germain et al., 2015, Laviolette et al., 2017]

$$\mathbb{P}(Z > 0.5) = \mathbb{P}(Z - \mathbb{E}[Z] \geq 0.5 - \mathbb{E}[Z])$$

by Chebyshev-Cantelli inequality, if $\mathbb{E}[Z] < 0.5$,

$$\leq \frac{\mathbb{V}[Z]}{(0.5 - \mathbb{E}[Z])^2 + \mathbb{V}[Z]}$$
$$= \frac{\mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2}{0.25 - \mathbb{E}[Z] + \mathbb{E}\left[Z^2\right]}$$

since $V[Z] = \mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2$.

Issues

- $\mathbb{E}\left[Z^2\right]$ and $\mathbb{E}[Z]$ in the denominator make empirical estimation and optimization of the bound difficult
- Empirically weaker than the first order bound [Lorenzen et al., 2019]

Tandem Bound (TND)[Masegosa et al., 2020]

Let $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$. By second order Markov's inequality $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{E}\left[Z^2\right]/\varepsilon^2$:

Tandem Bound (TND)[Masegosa et al., 2020]

Let $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$. By second order Markov's inequality $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{E}\left[Z^2\right]/\varepsilon^2$:

$$\begin{aligned} L(\mathrm{MV}_\rho) &\leq 4\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] \\ &= 4\,\underbrace{\mathbb{E}_{\rho^2}[L(h,h')]}_{\text{expected tandem loss}}. \end{aligned}$$

$\mathbb{E}_{\rho^2}[\cdot] = \mathbb{E}_{h\sim\rho,h'\sim\rho}[\cdot]$
$L(h,h') = \mathbb{E}_D[\ell(h,h')]$

Tandem Bound (TND)[Masegosa et al., 2020]

Let $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$. By second order Markov's inequality $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{E}\left[Z^2\right]/\varepsilon^2$:

$$\begin{aligned} L(\mathrm{MV}_\rho) &\leq 4\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] \\ &= 4\,\underbrace{\mathbb{E}_{\rho^2}[L(h, h')]}_{\text{expected tandem loss}}. \end{aligned}$$

$\mathbb{E}_{\rho^2}[\cdot] = \mathbb{E}_{h\sim\rho, h'\sim\rho}[\cdot]$
$L(h, h') = \mathbb{E}_D[\ell(h, h')]$

Tandem Loss

$$\ell(h, h') := \mathbb{1}\left(h(X) \neq Y \wedge h'(X) \neq Y\right)$$

counts an error only if h and h' both err on a sample (X, Y).

Tandem Bound (TND)[Masegosa et al., 2020]

Let $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$. By second order Markov's inequality $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{E}\left[Z^2\right]/\varepsilon^2$:

$$L(\mathsf{MV}_\rho) \leq 4\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] \\ = 4\underbrace{\mathbb{E}_{\rho^2}[L(h, h')]}_{\text{expected tandem loss}}.$$

$\mathbb{E}_{\rho^2}[\cdot] = \mathbb{E}_{h\sim\rho, h'\sim\rho}[\cdot]$
$L(h, h') = \mathbb{E}_D[\ell(h, h')]$

Tandem Loss

$$\ell(h, h') := \mathbb{1}\left(h(X) \neq Y \wedge h'(X) \neq Y\right)$$

counts an error only if h and h' both err on a sample (X, Y).

Issues

- Not as tight as the C-bound

State-of-the-art Summary

First order bound

- Ignores correlations
- Deteriorates test loss

State-of-the-art Summary

First order bound

- Ignores correlations
- Deteriorates test loss

C-bound

- Difficult to estimate and optimize due to $\mathbb{V}[Z]$ in denominator

State-of-the-art Summary

First order bound

- Ignores correlations
- Deteriorates test loss

C-bound

- Difficult to estimate and optimize due to $\mathbb{V}[Z]$ in denominator

Tandem bound

- Accounts for correlations
- Easy to optimize
- Not as tight as C-bound

State-of-the-art Summary

First order bound

- Ignores correlations
- Deteriorates test loss

C-bound

- Difficult to estimate and optimize due to $\mathbb{V}[Z]$ in denominator

Tandem bound

- Accounts for correlations
- Easy to optimize
- Not as tight as C-bound

Our contribution

- New form of Chebyshev-Cantelli inequality that has tightness of C-bound and is easier to optimization

Our Contribution (1)

New form of the Chebyshev-Cantelli inequality, which is convenient for optimization.

Our Contribution (1)

New form of the Chebyshev-Cantelli inequality, which is convenient for optimization.

Theorem (Parametrized Chebyshev-Cantelli inequality)

For any $\varepsilon > 0$ *and* $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon - \mu)^2}.$$

Our Contribution (1)

New form of the Chebyshev-Cantelli inequality, which is convenient for optimization.

Theorem (Parametrized Chebyshev-Cantelli inequality)

For any $\varepsilon > 0$ *and* $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2}.$$

Proof.

$$\begin{aligned}
&\mathbb{P}(Z \geq \varepsilon) \\
&= \mathbb{P}(Z-\mu \geq \varepsilon-\mu) \leq \mathbb{P}\left((Z-\mu)^2 \geq (\varepsilon-\mu)^2\right) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2}
\end{aligned}$$

$\square$

Relation to Existing Second Order Bounds

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

Relation to Existing Second Order Bounds

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

- Bound is minimized by $\mu^* = \mathbb{E}\left[Z\right] - \frac{\mathbb{V}[Z]}{\varepsilon - \mathbb{E}[Z]}$

Relation to Existing Second Order Bounds

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

- Bound is minimized by $\mu^* = \mathbb{E}\left[Z\right] - \frac{\mathbb{V}[Z]}{\varepsilon - \mathbb{E}[Z]}$
- Substitution of μ^* into the bound recovers Chebyshev-Cantelli inequality

Relation to Existing Second Order Bounds

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

- Bound is minimized by $\mu^* = \mathbb{E}\left[Z\right] - \frac{\mathbb{V}[Z]}{\varepsilon - \mathbb{E}[Z]}$
- Substitution of μ^* into the bound recovers Chebyshev-Cantelli inequality
- Taking $\mu = 0$ recovers second order Markov's inequality

Relation to Existing Second Order Bounds

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

- Bound is minimized by $\mu^* = \mathbb{E}\left[Z\right] - \frac{\mathbb{V}[Z]}{\varepsilon - \mathbb{E}[Z]}$
- Substitution of μ^* into the bound recovers Chebyshev-Cantelli inequality
- Taking $\mu = 0$ recovers second order Markov's inequality

Advantages

- Easy to estimate and optimize, as the second order Markov

Relation to Existing Second Order Bounds

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

- Bound is minimized by $\mu^* = \mathbb{E}\left[Z\right] - \frac{\mathbb{V}[Z]}{\varepsilon - \mathbb{E}[Z]}$
- Substitution of μ^* into the bound recovers Chebyshev-Cantelli inequality
- Taking $\mu = 0$ recovers second order Markov's inequality

Advantages

- Easy to estimate and optimize, as the second order Markov
- As tight as the Chebyshev-Cantelli inequality

Theorem 7

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

Theorem 7

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

Let $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$.

Theorem 7

In multiclass classification, if $\mu < 0.5$,

$$L(\mathrm{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L(h,h')] - 2\mu\mathbb{E}_\rho[L(h)] + \mu^2}{(0.5-\mu)^2}.$$

From Oracle to Empirical

In multiclass classification, if $\mu < 0.5$,

$$L(\mathsf{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L(h,h')] - 2\mu\mathbb{E}_\rho[L(h)] + \mu^2}{(0.5-\mu)^2}.$$

To bound $\mathbb{E}_{\rho^2}[L(h,h')]$ and $\mathbb{E}_\rho[L(h)]$ by their empirical counterparts, we use

From Oracle to Empirical

In multiclass classification, if $\mu < 0.5$,

$$L(\mathrm{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L(h, h')] - 2\mu\mathbb{E}_\rho[L(h)] + \mu^2}{(0.5 - \mu)^2}.$$

To bound $\mathbb{E}_{\rho^2}[L(h, h')]$ and $\mathbb{E}_\rho[L(h)]$ by their empirical counterparts, we use
PAC-Bayes-kl inequality [Seeger, 2002]:

$$\mathrm{kl}\left(\mathbb{E}_\rho[\hat{L}(h, S)] \middle\| \mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta)}{n}$$

From Oracle to Empirical

In multiclass classification, if $\mu < 0.5$,

$$L(\mathrm{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L(h,h')] - 2\mu\mathbb{E}_\rho[L(h)] + \mu^2}{(0.5-\mu)^2}.$$

To bound $\mathbb{E}_{\rho^2}[L(h,h')]$ and $\mathbb{E}_\rho[L(h)]$ by their empirical counterparts, we use
PAC-Bayes-λ inequality [Thiemann et al., 2016]:

$$\mathbb{E}_\rho\left[L(h)\right] \leq \frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta)}{\lambda\left(1-\frac{\lambda}{2}\right)n}$$

$$\mathbb{E}_\rho\left[L(h)\right] \geq \left(1-\frac{\gamma}{2}\right)\mathbb{E}_\rho[\hat{L}(h,S)] - \frac{\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta)}{\gamma n}$$

From Oracle to Empirical

In multiclass classification, if $\mu < 0.5$,

$$L(\mathrm{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L(h,h')] - 2\mu\mathbb{E}_\rho[L(h)] + \mu^2}{(0.5-\mu)^2}.$$

To bound $\mathbb{E}_{\rho^2}[L(h,h')]$ and $\mathbb{E}_\rho[L(h)]$ by their empirical counterparts, we use
PAC-Bayes-λ inequality [Thiemann et al., 2016]:

$$\mathbb{E}_\rho\left[L(h)\right] \leq \frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta)}{\lambda\left(1-\frac{\lambda}{2}\right)n}$$

$$\mathbb{E}_\rho\left[L(h)\right] \geq \left(1-\frac{\gamma}{2}\right)\mathbb{E}_\rho[\hat{L}(h,S)] - \frac{\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta)}{\gamma n}$$

$\Rightarrow$ **C**hebyshev-**C**antelli bound with **TND** empirical loss estimate

Theorem 8

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

Theorem 8

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

With $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ and $\mathbb{E}_{\rho^2}[\cdot]$ as a shorthand for $\mathbb{E}_{h\sim\rho, h'\sim\rho}[\cdot]$,

Theorem 8

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

With $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ and $\mathbb{E}_{\rho^2}[\cdot]$ as a shorthand for $\mathbb{E}_{h\sim\rho,h'\sim\rho}[\cdot]$,

$$\mathbb{E}\left[(Z-\mu)^2\right] = \mathbb{E}_D[(\mathbb{E}_\rho[(\mathbb{1}(h(X) \neq Y) - \mu)])^2]$$

Theorem 8

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

With $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ and $\mathbb{E}_{\rho^2}[\cdot]$ as a shorthand for $\mathbb{E}_{h\sim\rho, h'\sim\rho}[\cdot]$,

$$\begin{aligned}\mathbb{E}\left[(Z-\mu)^2\right] &= \mathbb{E}_D[(\mathbb{E}_\rho[(\mathbb{1}(h(X) \neq Y) - \mu)])^2] \\ &= \mathbb{E}_D[\mathbb{E}_{\rho^2}[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu)]]\end{aligned}$$

Theorem 8

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

With $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ and $\mathbb{E}_{\rho^2}[\cdot]$ as a shorthand for $\mathbb{E}_{h\sim\rho,h'\sim\rho}[\cdot]$,

$$\begin{aligned}
\mathbb{E}\left[(Z-\mu)^2\right] &= \mathbb{E}_D[(\mathbb{E}_\rho[(\mathbb{1}(h(X) \neq Y) - \mu)])^2] \\
&= \mathbb{E}_D[\mathbb{E}_{\rho^2}[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu)]] \\
&= \mathbb{E}_{\rho^2}[\mathbb{E}_D[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu)]]
\end{aligned}$$

Theorem 8

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

With $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ and $\mathbb{E}_{\rho^2}[\cdot]$ as a shorthand for $\mathbb{E}_{h\sim\rho,h'\sim\rho}[\cdot]$,

$$\begin{aligned}
\mathbb{E}\left[(Z-\mu)^2\right] &= \mathbb{E}_D[(\mathbb{E}_\rho[(\mathbb{1}(h(X) \neq Y) - \mu)])^2] \\
&= \mathbb{E}_D[\mathbb{E}_{\rho^2}[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu)]] \\
&= \mathbb{E}_{\rho^2}[\mathbb{E}_D[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu)]] \\
&= \underbrace{\mathbb{E}_{\rho^2}[L_\mu(h,h')]}_{\text{expected tandem loss with } \mu\text{-offset}} .
\end{aligned}$$

Theorem 8

For any $\varepsilon > 0$ and $\mu < \varepsilon$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}.$$

With $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ and $\mathbb{E}_{\rho^2}[\cdot]$ as a shorthand for $\mathbb{E}_{h\sim\rho,h'\sim\rho}[\cdot]$,

$$\begin{aligned}\mathbb{E}\left[(Z-\mu)^2\right] &= \mathbb{E}_D[(\mathbb{E}_\rho[(\mathbb{1}(h(X) \neq Y) - \mu)])^2] \\ &= \mathbb{E}_D[\mathbb{E}_{\rho^2}[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu)]] \\ &= \mathbb{E}_{\rho^2}[\mathbb{E}_D[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu)]] \\ &= \underbrace{\mathbb{E}_{\rho^2}[L_\mu(h, h')]}_{\text{expected tandem loss with } \mu\text{-offset}}.\end{aligned}$$

Tandem loss with μ-offset (μ-tandem loss)

$$\ell_\mu(h(X), h'(X), Y) = (\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu)$$

Theorem 8

In multiclass classification, if $\mu < 0.5$,

$$L(\mathsf{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L_\mu(h, h')]}{(0.5-\mu)^2}.$$

Theorem 8

In multiclass classification, if $\mu < 0.5$,

$$L(\mathsf{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L_\mu(h, h')]}{(0.5 - \mu)^2}.$$

μ-tandem loss

$$\ell_\mu(h(X), h'(X), Y) = (\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\big(h'(X) \neq Y\big) - \mu)$$

Theorem 8

In multiclass classification, if $\mu < 0.5$,

$$L(\mathsf{MV}_\rho) \le \frac{\mathbb{E}_{\rho^2}[L_\mu(h,h')]}{(0.5-\mu)^2}.$$

μ-tandem loss

$$\begin{aligned}\ell_\mu(h(X),h'(X),Y) &= (\mathbb{1}(h(X)\neq Y)-\mu)(\mathbb{1}\big(h'(X)\neq Y\big)-\mu)\\ &\in \left\{(1-\mu)^2, -\mu(1-\mu), \mu^2\right\}\end{aligned}$$

Theorem 8

In multiclass classification, if $\mu < 0.5$,

$$L(\mathsf{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L_\mu(h, h')]}{(0.5 - \mu)^2}.$$

μ-tandem loss

$$\begin{aligned}\ell_\mu(h(X), h'(X), Y) &= (\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}\left(h'(X) \neq Y\right) - \mu) \\ &\in \left\{(1-\mu)^2, -\mu(1-\mu), \mu^2\right\}\end{aligned}$$

Range $K_\mu = \max\{1 - \mu, 1 - 2\mu\}$.

Contribution (2): PAC-Bayes-Bennett Inequality

Contribution (2): PAC-Bayes-Bennett Inequality

Theorem (PAC-Bayes-Bernstein [Seldin et al., 2012] (Informal))

Assume $|\tilde{\ell}(\cdot,\cdot)| \leq b$ *and the corresponding variance is finite. Then for* $\gamma \in (0, 1/b]$,

$$\mathbb{E}_\rho[\tilde{L}(h)] \leq \mathbb{E}_\rho[\hat{\tilde{L}}(h,S)] + (e-2)\gamma\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\gamma n}.$$

Theorem (PAC-Bayes-Bennett (Informal))

Assume $\tilde{\ell}(\cdot,\cdot) \leq b$ *and the corresponding variance is finite. Let* $\phi(x) = e^x - x - 1$. *Then for* $\gamma > 0$,

$$\mathbb{E}_\rho[\tilde{L}(h)] \leq \mathbb{E}_\rho[\hat{\tilde{L}}(h,S)] + \frac{\phi(\gamma b)}{\gamma b^2}\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\gamma n}.$$

Contribution (2): PAC-Bayes-Bennett Inequality

Theorem (PAC-Bayes-Bernstein [Seldin et al., 2012] (Informal))

Assume $|\tilde{\ell}(\cdot,\cdot)| \leq b$ *and the corresponding variance is finite. Then for* $\gamma \in (0, 1/b]$,

$$\mathbb{E}_\rho[\tilde{L}(h)] \leq \mathbb{E}_\rho[\hat{\tilde{L}}(h,S)] + (e-2)\gamma\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\gamma n}.$$

Theorem (PAC-Bayes-Bennett (Informal))

Assume $\tilde{\ell}(\cdot,\cdot) \leq b$ *and the corresponding variance is finite. Let* $\phi(x) = e^x - x - 1$. *Then for* $\gamma > 0$,

$$\mathbb{E}_\rho[\tilde{L}(h)] \leq \mathbb{E}_\rho[\hat{\tilde{L}}(h,S)] + \frac{\phi(\gamma b)}{\gamma b^2}\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\gamma n}.$$

Note: $0.5 \leq \frac{\phi(\gamma b)}{\gamma^2 b^2} \leq (e-2) \approx 0.72$

Bound the Variance [Tolstikhin and Seldin, 2013]

Assume $\tilde{\ell}(\cdot,\cdot)$ has range c. For any $\lambda \in \left(0, \frac{2(n-1)}{n}\right)$,

$$\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)] \leq \frac{\mathbb{E}_\rho[\hat{\tilde{\mathbb{V}}}(h)]}{1 - \frac{\lambda n}{2(n-1)}} + \frac{c^2\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}\right)}{n\lambda\left(1 - \frac{\lambda n}{2(n-1)}\right)}.$$

From Oracle to Empirical

Parametrized Chebyshev-Cantelli oracle

If $\mu < 0.5$,

$$L(\mathrm{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L_\mu(h, h')]}{(0.5 - \mu)^2}.$$

Chebyshev-**C**antelli bound with **P**AC-**B**ayes-**B**ennett loss estimate

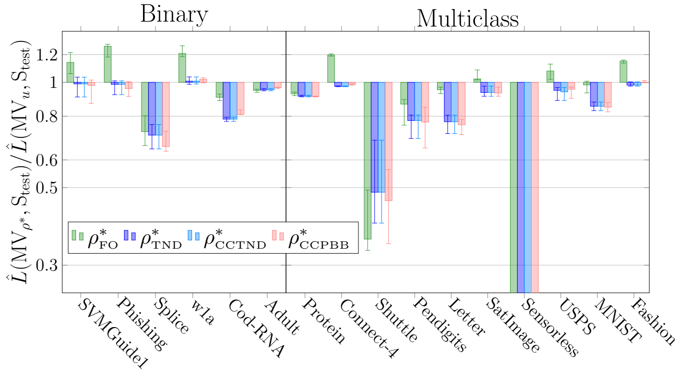
Theorem (Informal)

$$L(\mathrm{MV}_\rho) \leq \frac{1}{(0.5-\mu)^2}\left(\mathbb{E}_{\rho^2}[\hat{L}_\mu(h, h', S)] + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2k_\gamma k_\lambda}{\delta}}{\gamma n}\right.$$
$$\left. + \frac{\phi(\gamma(1-\mu)^2)}{\gamma(1-\mu)^4}\left(\frac{\mathbb{E}_{\rho^2}[\hat{\mathbb{V}}_\mu(h, h', S)]}{1 - \frac{\lambda n}{2(n-1)}} + \frac{K_\mu^2\left(2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2k_\gamma k_\lambda}{\delta}\right)}{n\lambda\left(1 - \frac{\lambda n}{2(n-1)}\right)}\right)\right).$$

k_γ, k_λ: # of parameter grid of γ and λ.

Experiment

Restrict $\mu \in [0, 0.5)$. Test error of optimized majority vote over uniformly weighted baseline for the first order bound, the TND bound and the two new bounds, CCTND and CCPBB. The lower the better.



Summary: Whats New?

Summary: Whats New?

- Parametric form of Chebyshev-Cantelli inequality

Summary: Whats New?

- Parametric form of Chebyshev-Cantelli inequality
  - No variance in the denominator and as tight as original bound
  - Allows efficient minimization and empirical estimation

Summary: Whats New?

- Parametric form of Chebyshev-Cantelli inequality
  - No variance in the denominator and as tight as original bound
  - Allows efficient minimization and empirical estimation
- New second order oracle bounds for weighted majority vote

Summary: Whats New?

- Parametric form of Chebyshev-Cantelli inequality
  - No variance in the denominator and as tight as original bound
  - Allows efficient minimization and empirical estimation
- New second order oracle bounds for weighted majority vote
  - Resulting empirical bounds are amenable to efficient minimization

Summary: Whats New?

- Parametric form of Chebyshev-Cantelli inequality
  - No variance in the denominator and as tight as original bound
  - Allows efficient minimization and empirical estimation
- New second order oracle bounds for weighted majority vote
  - Resulting empirical bounds are amenable to efficient minimization
- PAC-Bayes-Bennett inequality

Summary: Whats New?

- Parametric form of Chebyshev-Cantelli inequality
  - No variance in the denominator and as tight as original bound
  - Allows efficient minimization and empirical estimation
- New second order oracle bounds for weighted majority vote
  - Resulting empirical bounds are amenable to efficient minimization
- PAC-Bayes-Bennett inequality
  - Improves on the PAC-Bayes-Bernstein inequality by Seldin et al. [2012]
  - Can be used for bounding the tandem loss with an offset