Second Order PAC-Bayesian Bounds for the Weighted Majority Vote

Andrés R. Masegosa[1] Stephan S. Lorenzen[2]
Christian Igel[2] Yevgeny Seldin[2]

[1]University of Almería

[2]University of Copenhagen

NeurIPS, December 2020

Weighted Majority Vote

- Fundamental technique for combining predictions of multiple classifiers
- Used in Bagging, Boosting, etc.
- Wins most ML competitions

Weighted Majority Vote

- Fundamental technique for combining predictions of multiple classifiers
- Used in Bagging, Boosting, etc.
- Wins most ML competitions

Ensemble's Key Power

- **Cancellation of errors effect**
  If the errors are independent, they average out

Weighted Majority Vote

- Fundamental technique for combining predictions of multiple classifiers
- Used in Bagging, Boosting, etc.
- Wins most ML competitions

Ensemble's Key Power

- **Cancellation of errors effect**
  If the errors are independent, they average out

Our contributions

- Second order PAC-Bayesian generalization bound for the weighted majority vote
- Minimization of the bound guides weighting of ensemble members and does not deteriorate the test error

Standard analysis

If ρ-weighted majority vote makes an error, then at least a ρ-weighted half of the classifiers make an error

Standard analysis

If ρ-weighted majority vote makes an error, then at least a ρ-weighted half of the classifiers make an error

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \quad \leq \quad \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass} \\ \text{of errors}}} \geq 0.5 \Big)$$

Standard analysis

If ρ-weighted majority vote makes an error, then at least a ρ-weighted half of the classifiers make an error

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \leq \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass} \\ \text{of errors}}} \geq 0.5 \Big)$$

First order bound: use Markov's inequality ($\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}[X]$)

$$\begin{aligned} &\leq 2\, \mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] \\ &= 2\, \underbrace{\mathbb{E}_\rho[L(h)]}_{\substack{\text{Expected loss of } \rho\text{-weighted} \\ \text{randomized classifier}}} \end{aligned}$$

Standard analysis

If ρ-weighted majority vote makes an error, then at least a ρ-weighted half of the classifiers make an error

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \leq \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass} \\ \text{of errors}}} \geq 0.5 \Big)$$

First order bound: use Markov's inequality ($\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}[X]$)

$$\begin{aligned} &\leq 2\, \mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] \\ &= 2\, \underbrace{\mathbb{E}_\rho[L(h)]}_{\substack{\text{Expected loss of } \rho\text{-weighted} \\ \text{randomized classifier}}} \end{aligned}$$

Issues

- Ignores correlation of errors (the key power)
- Minimization of the corresponding PAC-Bayes bound degrades the test error (Lorenzen et al., 2019)

Prior second order analysis

The C-bounds (Lacasse et al., 2007, Germain et al., 2015, Laviolette et al., 2017)

Based on Chebyshev-Cantelli inequality

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

Prior second order analysis

The C-bounds (Lacasse et al., 2007, Germain et al., 2015, Laviolette et al., 2017)

Based on Chebyshev-Cantelli inequality

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

Issues

- $\mathbb{E}\left[X^2\right]$ and $\mathbb{E}\left[X\right]$ in the denominator make empirical estimation hard
- Empirically weaker than the first order bound (Lorenzen et al., 2019)
- Impossible to optimize the weighting except in very restrictive cases

Our contribution: a novel second-order oracle bound

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \quad \leq \quad \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass} \\ \text{of errors}}} \geq 0.5 \Big)$$

Our contribution: a novel second-order oracle bound

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \quad \leq \quad \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass} \\ \text{of errors}}} \geq 0.5 \Big)$$

Second-order Markov's inequality $\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left[X^2\right]$:

Our contribution: a novel second-order oracle bound

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \leq \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass} \\ \text{of errors}}} \geq 0.5 \Big)$$

Second-order Markov's inequality $\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}[X^2]$:

$$\leq 4\, \mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2]$$

Our contribution: a novel second-order oracle bound

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of}\\ \rho\text{-weighted majority vote}}} \leq \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass}\\ \text{of errors}}} \geq 0.5\Big)$$

Second-order Markov's inequality $\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}[X^2]$:

$$\leq 4\,\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2]$$

$$= 4\,\mathbb{E}_{\rho^2}[\underbrace{\mathbb{E}_D[\mathbb{1}\big(h(X) \neq Y \wedge h'(X) \neq Y\big)]}_{\text{Expected } \textbf{Tandem Loss}:\ L(h,h')}]$$

Our contribution: a novel second-order oracle bound

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \leq \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass} \\ \text{of errors}}} \geq 0.5 \Big)$$

Second-order Markov's inequality $\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}[X^2]$:

$$\begin{aligned} &\leq 4\, \mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] \\ &= 4\, \mathbb{E}_{\rho^2}[\underbrace{\mathbb{E}_D[\mathbb{1}\big(h(X) \neq Y \wedge h'(X) \neq Y\big)]}_{\text{Expected } \textbf{Tandem Loss}\text{: } L(h,h')}] \\ &= 4\, \mathbb{E}_{\rho^2}[L(h, h')] \end{aligned}$$

Our contribution: a novel second-order oracle bound

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of}\\ \rho\text{-weighted majority vote}}} \leq \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass}\\ \text{of errors}}} \geq 0.5\Big)$$

Second-order Markov's inequality $\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}[X^2]$:

$$\begin{aligned} &\leq 4\,\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] \\ &= 4\,\mathbb{E}_{\rho^2}[\underbrace{\mathbb{E}_D[\mathbb{1}\big(h(X) \neq Y \wedge h'(X) \neq Y\big)]}_{\text{Expected } \textbf{Tandem Loss}\text{: } L(h,h')}] \\ &= 4\,\mathbb{E}_{\rho^2}[L(h,h')] \end{aligned}$$

- **Tandem loss** counts an error if both h and h' err on a sample

Our contribution: a novel second-order oracle bound

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \leq \mathbb{P}\Big(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho\text{-weighted mass} \\ \text{of errors}}} \geq 0.5 \Big)$$

Second-order Markov's inequality $\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}[X^2]$:

$$\begin{aligned} &\leq 4\, \mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] \\ &= 4\, \mathbb{E}_{\rho^2}[\underbrace{\mathbb{E}_D[\mathbb{1}\big(h(X) \neq Y \wedge h'(X) \neq Y\big)]}_{\text{Expected } \textbf{Tandem Loss}\text{: } L(h,h')}] \\ &= 4\, \mathbb{E}_{\rho^2}[L(h, h')] \end{aligned}$$

- **Tandem loss** counts an error if both h and h' err on a sample
- **Second order oracle bound:** $L(\mathrm{MV}_\rho) \leq 4\mathbb{E}_{\rho^2}[L(h, h')]$

A specialized oracle bound for binary classification

In binary classification **tandem loss** $L(h, h')$ satisfies

$$\mathbb{E}_{\rho^2}[L(h,h')] = \underbrace{\mathbb{E}_{\rho}[L(h)]}_{\substack{\text{Expected loss of } \rho\text{-weighted} \\ \text{randomized classifier}}} - \frac{1}{2}\,\mathbb{E}_{\rho^2}[\underbrace{\mathbb{E}_{D}[\mathbb{1}\big(h(X) \neq h'(X)\big)]}_{\substack{\text{Expected Disagreement} \\ \mathbb{D}(h,h')}}].$$

A specialized oracle bound for binary classification

In binary classification **tandem loss** $L(h, h')$ satisfies

$$\mathbb{E}_{\rho^2}[L(h,h')] = \underbrace{\mathbb{E}_{\rho}[L(h)]}_{\substack{\text{Expected loss of } \rho\text{-weighted} \\ \text{randomized classifier}}} - \frac{1}{2}\,\mathbb{E}_{\rho^2}[\underbrace{\mathbb{E}_{D}[\mathbb{1}\big(h(X) \neq h'(X)\big)]}_{\substack{\text{Expected Disagreement} \\ \mathbb{D}(h,h')}}].$$

Specialized oracle bound for binary classification

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \leq 4\,\underbrace{\mathbb{E}_{\rho}[L(h)]}_{\substack{\text{Expected loss of } \rho\text{-weighted} \\ \text{randomized classifier}}} - 2\,\underbrace{\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]}_{\substack{\text{Expected Disagreement of} \\ \rho\text{-weighted rand. classifier}}}$$

A specialized oracle bound for binary classification

In binary classification **tandem loss** $L(h, h')$ satisfies

$$\mathbb{E}_{\rho^2}[L(h,h')] = \underbrace{\mathbb{E}_{\rho}[L(h)]}_{\substack{\text{Expected loss of } \rho\text{-weighted} \\ \text{randomized classifier}}} - \frac{1}{2}\,\mathbb{E}_{\rho^2}[\underbrace{\mathbb{E}_{D}[\mathbb{1}\left(h(X) \neq h'(X)\right)]}_{\substack{\text{Expected Disagreement} \\ \mathbb{D}(h,h')}}].$$

Specialized oracle bound for binary classification

$$\underbrace{L(\mathrm{MV}_\rho)}_{\substack{\text{Expected loss of} \\ \rho\text{-weighted majority vote}}} \leq 4\,\underbrace{\mathbb{E}_{\rho}[L(h)]}_{\substack{\text{Expected loss of } \rho\text{-weighted} \\ \text{randomized classifier}}} - 2\,\underbrace{\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]}_{\substack{\text{Expected Disagreement of} \\ \rho\text{-weighted rand. classifier}}}$$

$\mathbb{D}(h, h')$ only depends on **unlabeled data**!!

From oracle to empirical bounds

PAC-Bayes-λ (Thiemann et al., 2017):

For π independent of S, with probability at least $1 - \delta$ for all ρ and $\lambda \in (0, 2)$

From oracle to empirical bounds

PAC-Bayes-λ (Thiemann et al., 2017):

For π independent of S, with probability at least $1-\delta$ for all ρ and $\lambda \in (0, 2)$

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

From oracle to empirical bounds

PAC-Bayes-λ (Thiemann et al., 2017):

For π independent of S, with probability at least $1-\delta$ for all ρ and $\lambda \in (0,2)$ and $\gamma > 0$

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

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

Second-order PAC-Bayesian bound

$$L(\mathrm{MV}_\rho) \leq 4 \overbrace{\mathbb{E}_{\rho^2}[L(h, h')]}^{\substack{\text{Expected} \\ \textbf{Tandem Loss}}}$$

Second-order PAC-Bayesian bound

$$L(\mathrm{MV}_\rho) \leq 4 \overbrace{\mathbb{E}_{\rho^2}[L(h,h')]}^{\substack{\text{Expected} \\ \textbf{Tandem Loss}}}$$

$$\leq 4 \Bigg(\overbrace{\frac{\mathbb{E}_{\rho^2}[\hat{L}(h,h',S)]}{1-\lambda/2}}^{\substack{\text{Empirical} \\ \textbf{Tandem loss}}} + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n\lambda(1-\lambda/2)} \Bigg)$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{PAC-Bayes upper bound on } \mathbb{E}_{\rho^2}[L(h,h')]}$$

Second-order PAC-Bayesian bound

$$L(\mathrm{MV}_\rho) \leq 4 \overbrace{\mathbb{E}_{\rho^2}[L(h,h')]}^{\substack{\text{Expected} \\ \textbf{Tandem Loss}}}$$

$$\leq 4 \Bigg(\overbrace{\frac{\mathbb{E}_{\rho^2}[\hat{L}(h,h',S)]}{1-\lambda/2}}^{\substack{\text{Empirical} \\ \textbf{Tandem loss}}} + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n\lambda(1-\lambda/2)} \Bigg)$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\text{PAC-Bayes upper bound on } \mathbb{E}_{\rho^2}[L(h,h')]}$$

Advantages

- Takes correlation of errors into account
- Easy to minimize and tight
- Minimization of the bound does not degrade the test error

Second-order PAC-Bayesian bound for **binary classification**

$$\overbrace{L(\mathrm{MV}_\rho)}^{\substack{\text{Expected loss of}\\ \text{majority vote}}} \leq 4\,\mathbb{E}_\rho[L(h)] - 2\,\overbrace{\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]}^{\text{Expected Disagreement}}$$

Second-order PAC-Bayesian bound for **binary classification**

$$\overbrace{L(\mathrm{MV}_\rho)}^{\substack{\text{Expected loss of}\\ \text{majority vote}}} \leq 4\,\mathbb{E}_\rho[L(h)] - 2\,\overbrace{\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]}^{\text{Expected Disagreement}}$$

$$\leq 4\Bigg(\underbrace{\frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\lambda/2} + \frac{\mathrm{KL}(\rho\|\pi) + \ln(4\sqrt{n}/\delta)}{n\lambda(1-\lambda/2)}}_{\text{PAC-Bayes } \textbf{upper} \text{ bound on } \mathbb{E}_{\rho^2}[L(h,h')]}\Bigg)$$

Second-order PAC-Bayesian bound for **binary classification**

$$\overbrace{L(\mathrm{MV}_\rho)}^{\text{Expected loss of majority vote}} \leq 4\,\mathbb{E}_\rho[L(h)] - 2\,\overbrace{\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]}^{\text{Expected Disagreement}}$$

$$\leq 4\Bigg(\underbrace{\frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\lambda/2} + \frac{\mathrm{KL}(\rho\|\pi) + \ln(4\sqrt{n}/\delta)}{n\lambda(1-\lambda/2)}}_{\text{PAC-Bayes \textbf{upper} bound on } \mathbb{E}_{\rho^2}[L(h,h')]}\Bigg)$$

$$- 2\Bigg(\underbrace{(1-\gamma/2)\,\overbrace{\mathbb{E}_{\rho^2}[\hat{\mathbb{D}}(h,h',S')]}^{\text{Empirical Disagreement}} - \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln(4\sqrt{m}/\delta)}{\gamma m}}_{\text{PAC-Bayes \textbf{lower} bound on } \mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]}\Bigg)$$

Second-order PAC-Bayesian bound for **binary classification**

$$\overbrace{L(\mathrm{MV}_\rho)}^{\text{Expected loss of majority vote}} \leq 4\,\mathbb{E}_\rho[L(h)] - 2\,\overbrace{\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]}^{\text{Expected Disagreement}}$$

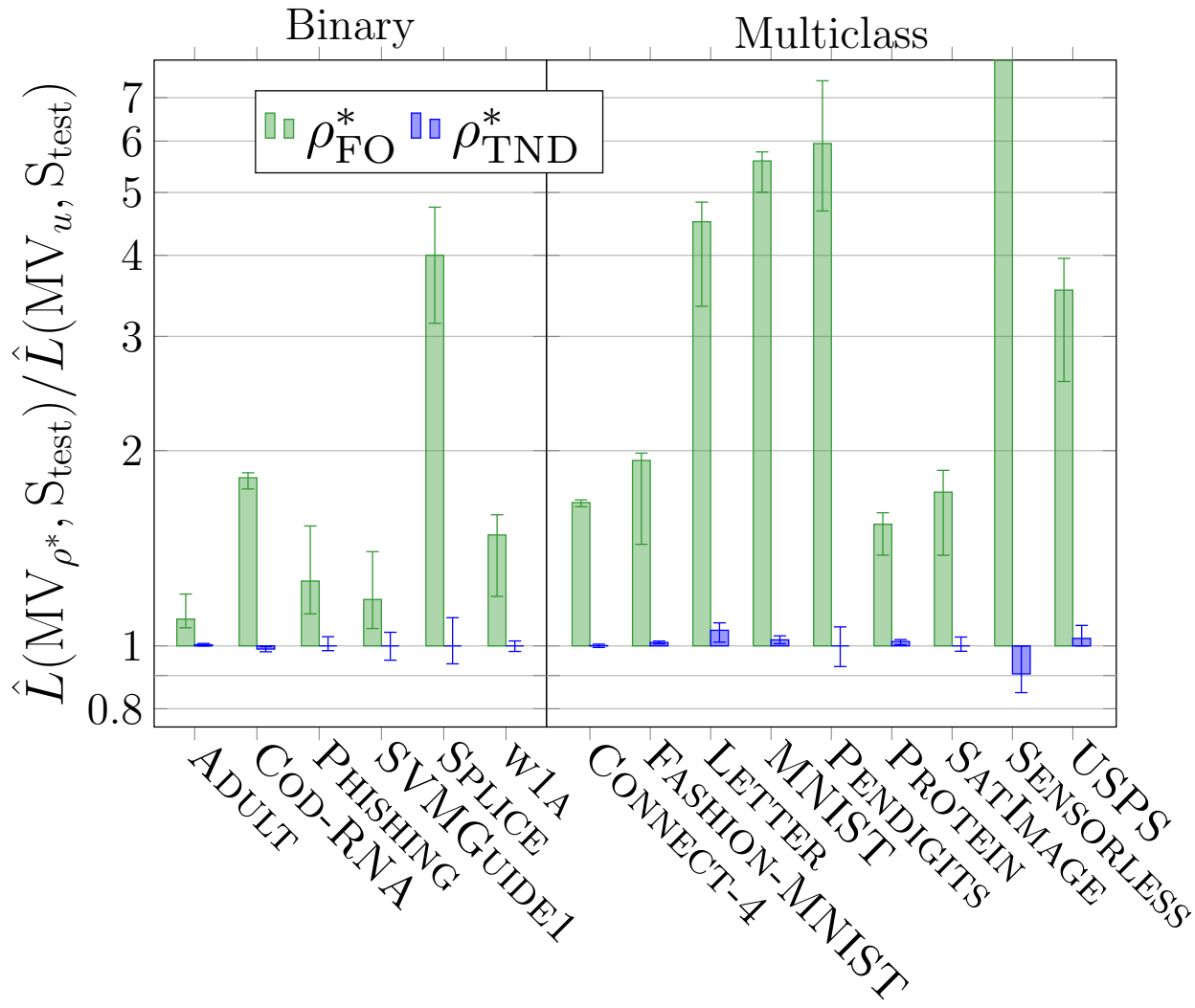
$$\leq 4\Bigg(\underbrace{\frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\lambda/2} + \frac{\mathrm{KL}(\rho\|\pi) + \ln(4\sqrt{n}/\delta)}{n\lambda(1-\lambda/2)}}_{\text{PAC-Bayes \textbf{upper} bound on } \mathbb{E}_{\rho^2}[L(h,h')]}\Bigg)$$

$$- 2\Bigg(\underbrace{(1-\gamma/2)\overbrace{\mathbb{E}_{\rho^2}[\hat{\mathbb{D}}(h,h',S')]}^{\text{Empirical Disagreement}} - \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln(4\sqrt{m}/\delta)}{\gamma m}}_{\text{PAC-Bayes \textbf{lower} bound on } \mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]}\Bigg)$$

- It can exploit unlabeled data

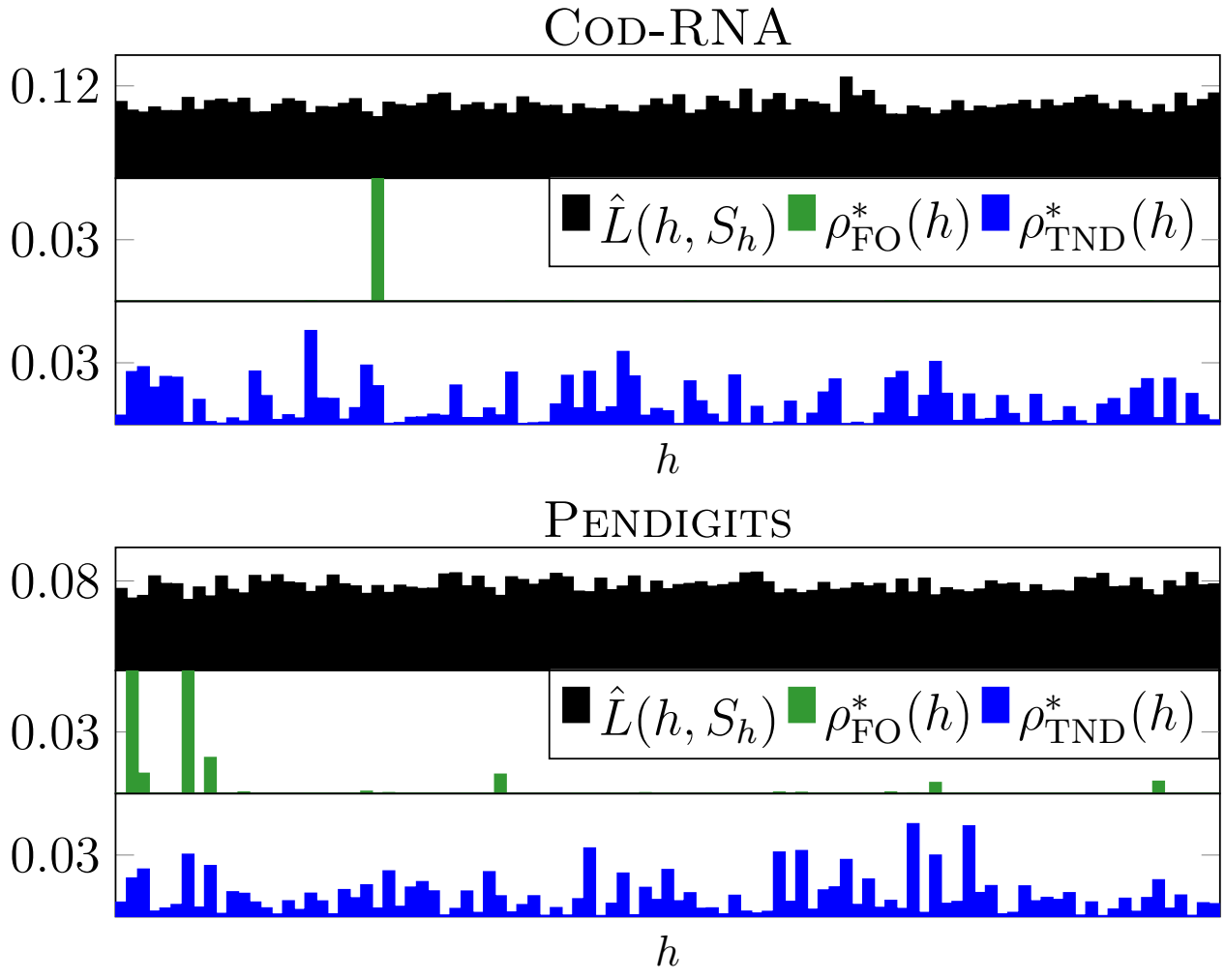
Empirical evaluation

- Test error of optimized majority vote over uniformly weighted baseline for first order [FO] and new second order [TND] bound (the lower the better)



Empirical evaluation

- The optimized weights $\rho^\star$ generated by the first order [FO] and the new second order [TND] bound.



Summary

State-of-the-art

- Minimization of existing first-order bound significantly deteriorates the test error

Summary

State-of-the-art

- Minimization of existing first-order bound significantly deteriorates the test error
- Existing second-order bounds are looser and can not be optimized

Summary

State-of-the-art

- Minimization of existing first-order bound significantly deteriorates the test error
- Existing second-order bounds are looser and can not be optimized

Contributions

- Novel second order oracle bound for the weighted majority vote based on second order Markov's inequality
- Novel second order PAC-Bayesian bound for the weighted majority vote

Summary

State-of-the-art

- Minimization of existing first-order bound significantly deteriorates the test error
- Existing second-order bounds are looser and can not be optimized

Contributions

- Novel second order oracle bound for the weighted majority vote based on second order Markov's inequality
- Novel second order PAC-Bayesian bound for the weighted majority vote
- Minimization of the bound guides weighting of ensemble members and does not deteriorate the test error