Fairness via Loss Variance Regularization



Fereshte Khani Aditi Raghunathan Percy Liang

EETimes

Microsoft, Google Beat Humans at Image Recognition

Deep learning algorithms compete at ImageNet challenge

By R. Colin Johnson, 02.18.15 14

EETimes

DESIGNLINES | INDUSTRIAL CONTROL DESIGNLINE

Microsoft, Google Beat Humans at Image Recognition

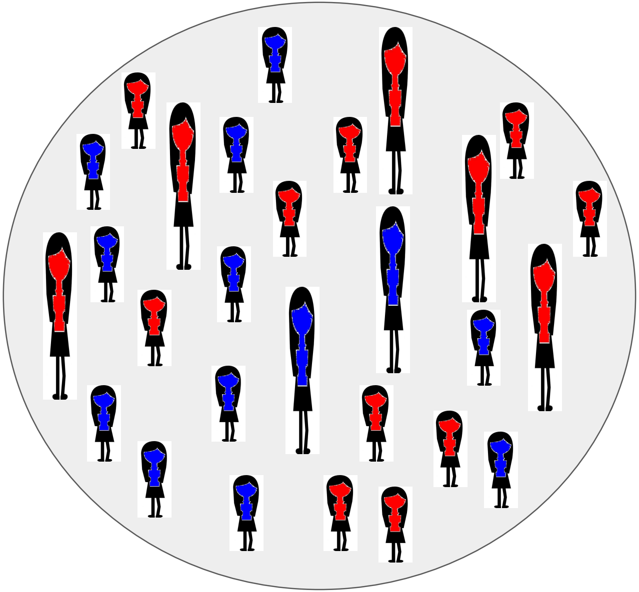
Deep learning algorithms compete at ImageNet challenge

By R. Colin Johnson, 02.18.15 14

The New York Times

Facial Recognition Is Accurate, if You're a White Guy

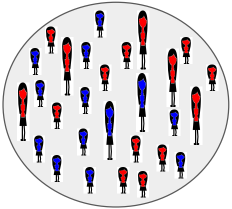
By Steve Lohr



$$\mathbb{E}[\ell] = 0.3$$

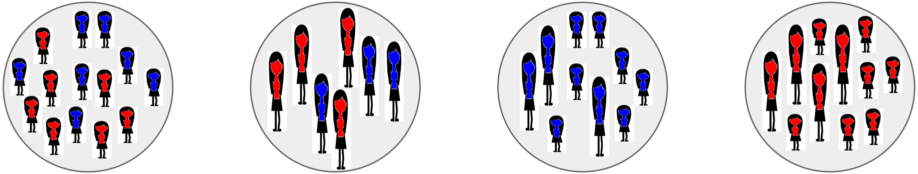
Sensitive attributes

A = [Height, Color]



$$\mathbb{E}[\ell] = 0.3$$

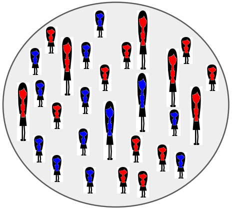
short tall blue red



Calders et al. [2009]

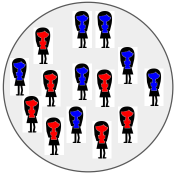
Sensitive attributes

A = [Height, Color]



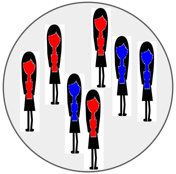
$\mathbb{E}[\ell] = 0.3$

short



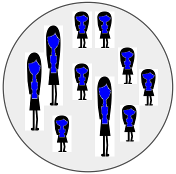
$\mathbb{E}[\ell] = 0.3$

tall



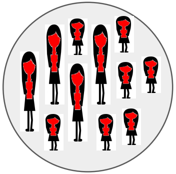
$\mathbb{E}[\ell] = 0.3$

blue



$\mathbb{E}[\ell] = 0.3$

red

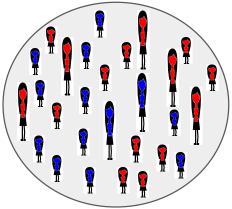


$\mathbb{E}[\ell] = 0.3$

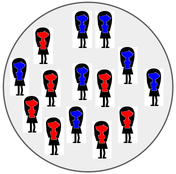
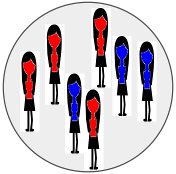
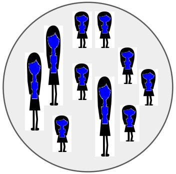
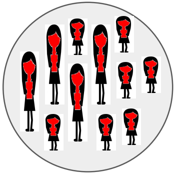
Calders et al. [2009]

Sensitive attributes

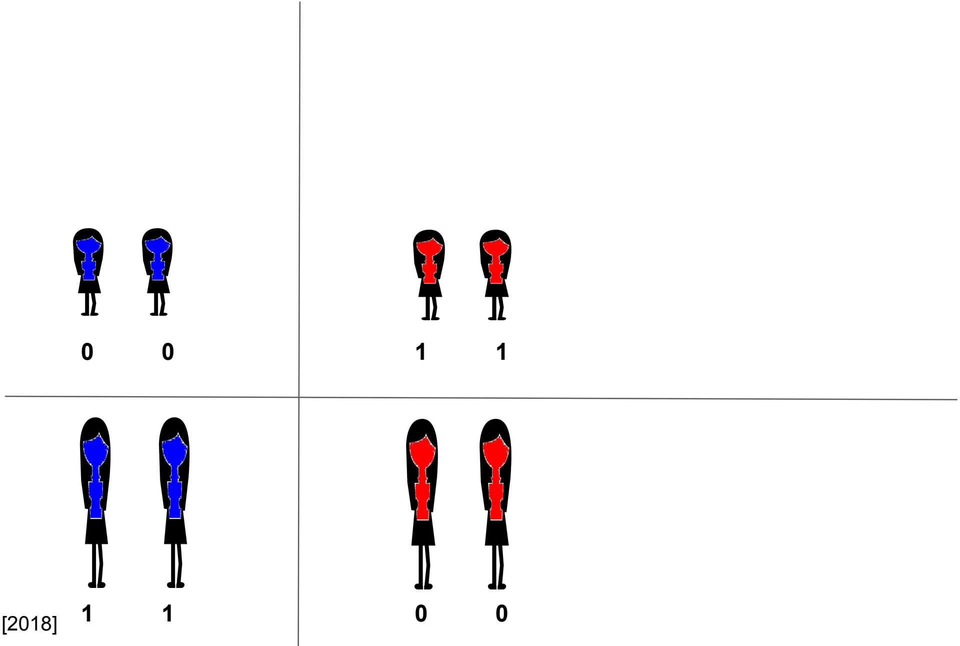
A = [Height, Color]



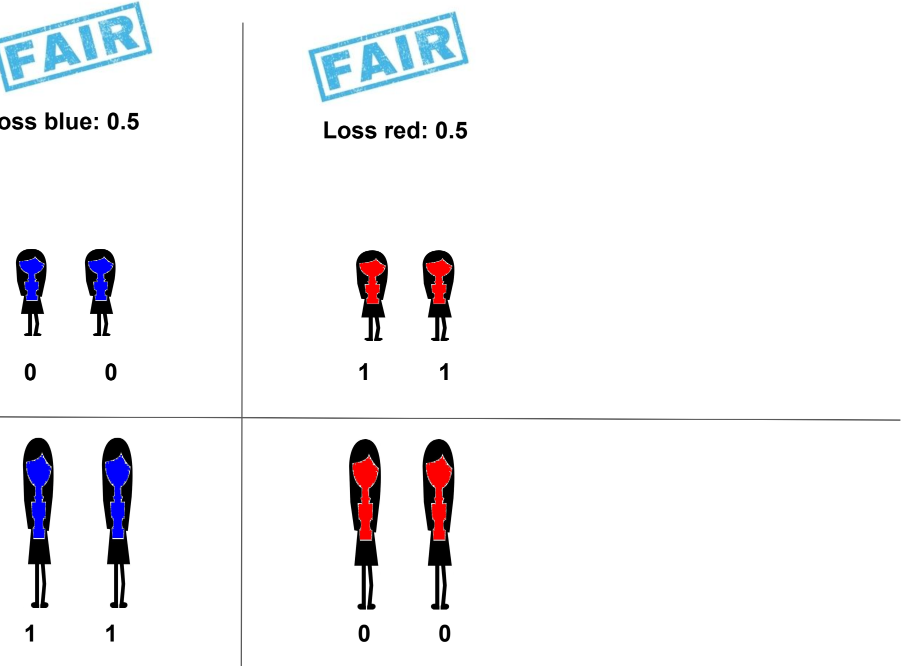
$\mathbb{E}[\ell] = 0.3$

	short	tall	blue	red
				
FAIR	$\mathbb{E}[\ell] = 0.3$	$\mathbb{E}[\ell] = 0.3$	$\mathbb{E}[\ell] = 0.3$	$\mathbb{E}[\ell] = 0.3$
UNFAIR!	$\mathbb{E}[\ell] = 0$	$\mathbb{E}[\ell] = 0.5$	$\mathbb{E}[\ell] = 0$	$\mathbb{E}[\ell] = 0.5$

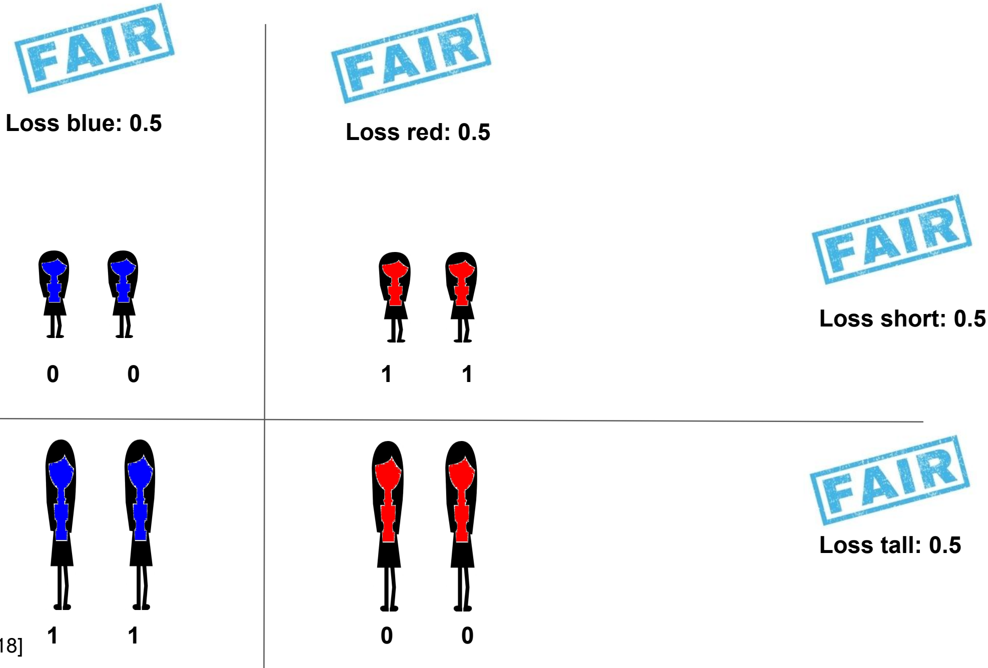
Calders et al. [2009]



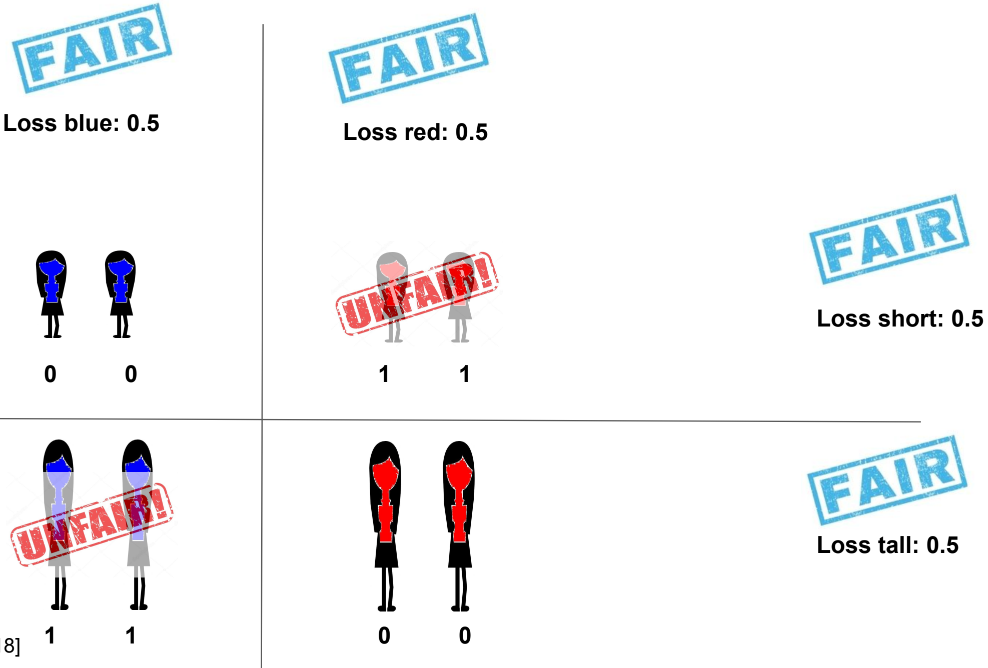
Kearns et al. [2018]



Kearns et al. [2018]

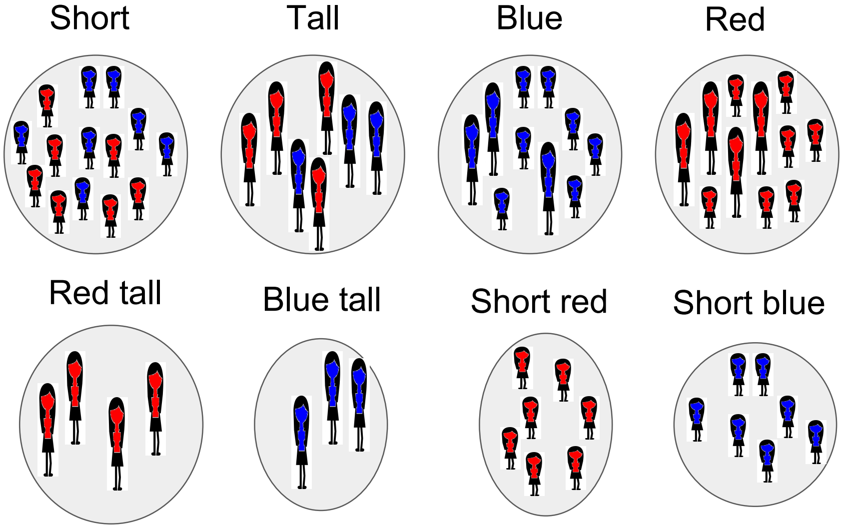


Kearns et al. [2018]



Kearns et al. [2018]

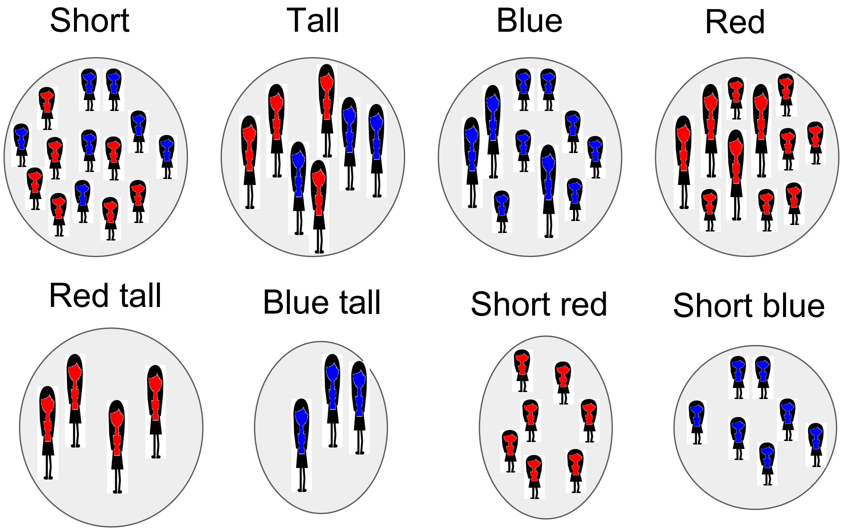
G_{sub}



Kearns et al. [2018]

$$G_{sub}$$

$$VC(G_{sub}) \leq C$$

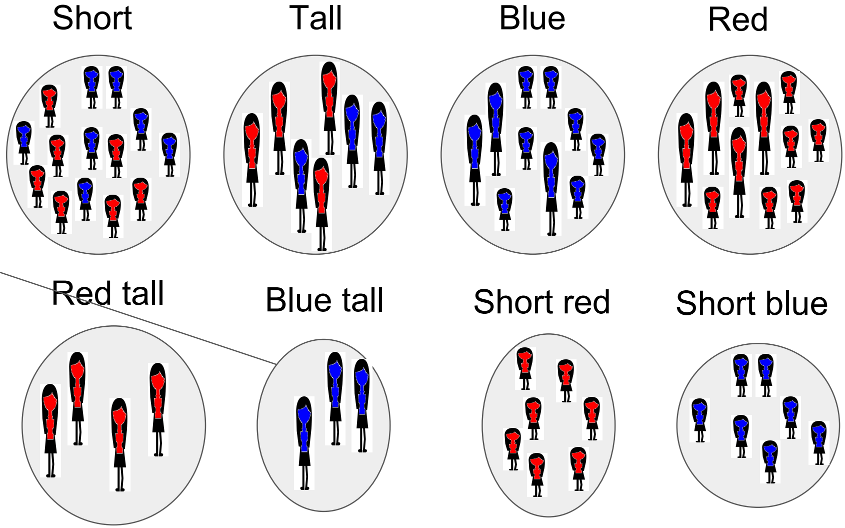


Kearns et al. [2018]

$$G_{sub}$$

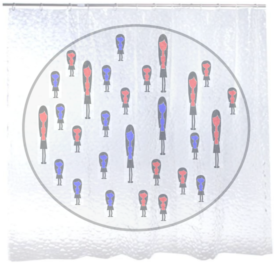
$$VC(G_{sub}) \leq C$$

Smaller groups can have higher loss



Kearns et al. [2018]

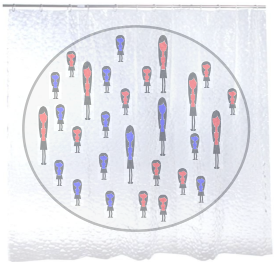
No sensitive attributes is available
A = [?]



$$\mathbb{E}[\ell] = 0.3$$

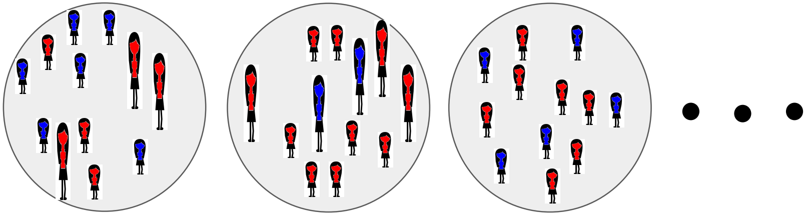
Hashimoto et al. 2018

No sensitive attributes is available
A = [?]



$\mathbb{E}[\ell] = 0.3$

All groups larger than α



Hashimoto et al. 2018

Maximum weighted loss discrepancy

Group weight

Group loss

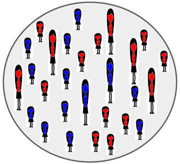
population loss

$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell] \right|$$

All possible groups on population

$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell] \right|$$

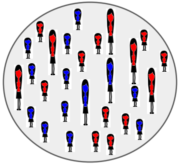
Normal



$$w(g) = \begin{cases} 1 & g = \text{all points} \\ 0 & o.w. \end{cases}$$

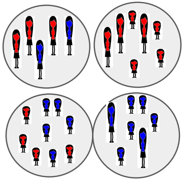
$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell] \right|$$

Normal



$$w(g) = \begin{cases} 1 & g = \text{all points} \\ 0 & o.w. \end{cases}$$

Group fairness

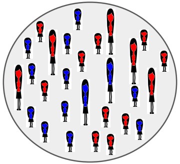


Hardt et al. 2016

$$w(g) = \begin{cases} 1 & sensitive \\ 0 & o.w. \end{cases}$$

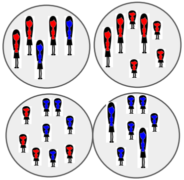
$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell] \right|$$

Normal



$$w(g) = \begin{cases} 1 & g = \text{all points} \\ 0 & o.w. \end{cases}$$

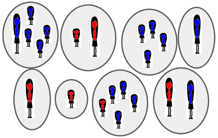
Group fairness



Hardt et al. 2016

$$w(g) = \begin{cases} 1 & sensitive \\ 0 & o.w. \end{cases}$$

Subgroup fairness

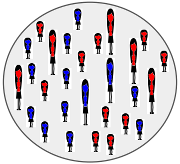


kearns et al. 2018

$$w(g) = \begin{cases} \mathbb{E}[g] & g \in G_{sub} \\ 0 & o.w. \end{cases}$$

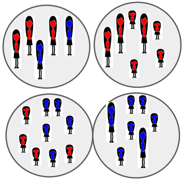
$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell] \right|$$

Normal



$$w(g) = \begin{cases} 1 & g = \text{all points} \\ 0 & o.w. \end{cases}$$

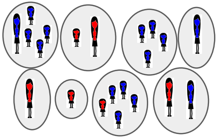
Group fairness



Hardt et al. 2016

$$w(g) = \begin{cases} 1 & sensitive \\ 0 & o.w. \end{cases}$$

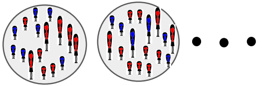
Subgroup fairness



kearns et al. 2018

$$w(g) = \begin{cases} \mathbb{E}[g] & g \in G_{sub} \\ 0 & o.w. \end{cases}$$

Large-group fairness



Hashimoto et al. 2018

$$w(g) = \begin{cases} 1 & \mathbb{E}[g] \geq \alpha \\ 0 & o.w. \end{cases}$$

Interpretation

$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g] - \mathbb{E}[\ell] \right|$$

Interpretation

$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g] - \mathbb{E}[\ell] \right|$$

For any g, we have:

$$\mathbb{E}[\ell] - \frac{U(w)}{w(g)} \leq \mathbb{E}[\ell \mid g] \leq \mathbb{E}[\ell] + \frac{U(w)}{w(g)}$$

Interpretation

$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g] - \mathbb{E}[\ell] \right|$$

For any g, we have:

$$\mathbb{E}[\ell] - \frac{U(w)}{w(g)} \leq \mathbb{E}[\ell \mid g] \leq \mathbb{E}[\ell] + \frac{U(w)}{w(g)}$$

Note. If $w(g) = 0$ then $\mathbb{E}[\ell \mid g]$ can be arbitrary even when $U(w)$ is small.

Without any prior information, what is an appropriate weighting function?

Without any prior information, what is an appropriate weighting function?

$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g] - \mathbb{E}[\ell] \right|$$

$$w(g) = ?$$

$$w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$$

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Proposition

There is no auditor for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Proposition

There is no auditor for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

Proof idea.

- Let $\mathcal{P}_1$ be any distribution such that $U(w) < \frac{1}{2}$ for $\mathcal{P}_1$.

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Proposition

There is no auditor for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

Proof idea.

- Let $\mathcal{P}_1$ be any distribution such that $U(w) < \frac{1}{2}$ for $\mathcal{P}_1$.
- Construct $\mathcal{P}_2$ such that $z \sim \mathcal{P}_1$ with probability $1 - \eta$ and $z = z_0$ and $z = z_1$ each with probability $\frac{\eta}{2} \implies U(w) \geq \frac{1}{2}$ for $\mathcal{P}_2$

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Proposition

There is no auditor for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

Proof idea.

- Let $\mathcal{P}_1$ be any distribution such that $U(w) < \frac{1}{2}$ for $\mathcal{P}_1$.
- Construct $\mathcal{P}_2$ such that $z \sim \mathcal{P}_1$ with probability $1 - \eta$ and $z = z_0$ and $z = z_1$ each with probability $\frac{\eta}{2} \implies U(w) \geq \frac{1}{2}$ for $\mathcal{P}_2$
- Choose a small $\eta = 1 - \sqrt[n]{2\delta}$.

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Proposition

There is no auditor for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

Take aways.

If we want to have positive weights for all groups then smaller groups should have smaller weights $\square$

Without any prior information, what is an appropriate weighting function?

Without any prior information, what is an appropriate weighting function?

$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g] - \mathbb{E}[\ell] \right|$$

$$w(g) = ?$$

$$w(g) = \mathbb{E}[g]^k \qquad k \in (0, 1]$$

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Theorem

Fix a bounded measurable loss. There is an auditor when $w(g) = \mathbb{E}[g]^k$ for $k \in (0, 1)$ which needs $n = \mathcal{O}\left(\frac{\ln(1/\delta)}{\epsilon^{2+\frac{1}{k}}}\right)$ data points.

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Theorem

Fix a bounded measurable loss. There is an auditor when $w(g) = \mathbb{E}[g]^k$ for $k \in (0, 1)$ which needs $n = \mathcal{O}\left(\frac{\ln(1/\delta)}{\epsilon^{2+\frac{1}{k}}}\right)$ data points.

Example

$$k = \frac{1}{2} \qquad\qquad n = \mathcal{O}\left(\frac{\ln(1/\delta)}{\epsilon^4}\right)$$

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Theorem

Fix a bounded measurable loss. There is an auditor when $w(g) = \mathbb{E}[g]^k$ for $k \in (0, 1)$ which needs $n = \mathcal{O}\left(\frac{\ln(1/\delta)}{\epsilon^{2+\frac{1}{k}}}\right)$ data points.

Take aways.

For small k, we need more examples for convergence! □

Definition (Auditor)

Given any $\epsilon, \delta \in (0, \frac{1}{2})$ an auditor returns an estimate γ for $U(w)$ such that:

$$\mathbb{P}\left[|U(w) - \gamma| \leq \epsilon\right] \geq 1 - \delta$$

Theorem

Fix a bounded measurable loss. There is an auditor when $w(g) = \mathbb{E}[g]^k$ for $k \in (0, 1)$ which needs $n = \mathcal{O}\left(\frac{\ln(1/\delta)}{\epsilon^{2+\frac{1}{k}}}\right)$ data points.

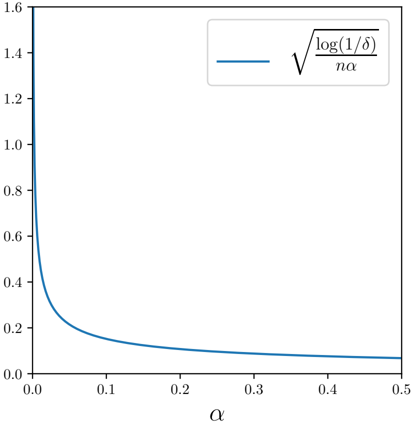
Proof idea.

- $\mathbb{P}\left[\left|U(w) - \hat{U}(w)\right| \leq \epsilon\right] \geq 1 - \delta$
- We can compute $\hat{U}(w)$ efficiently.

$\square$

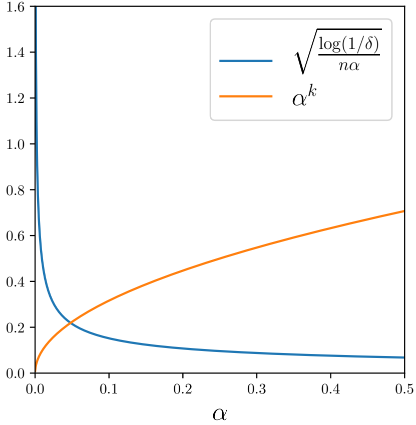
$D(g^\star)$ converges to $\widehat{D}(g^\star)$

- $D(g) = \mathbb{E}[g]^k \, |\mathbb{E}[\ell \mid g] - \mathbb{E}[\ell]|$
  (Weighted loss discrepancy)
- $g^\star = \arg\max_{g \in \mathcal{G}} D(g)$
- $\alpha = \mathbb{E}[g^\star]$
- $D(g^\star) - \widehat{D}(g^\star) \leq \sqrt{\frac{\ln(1/\delta)}{n\alpha}}$



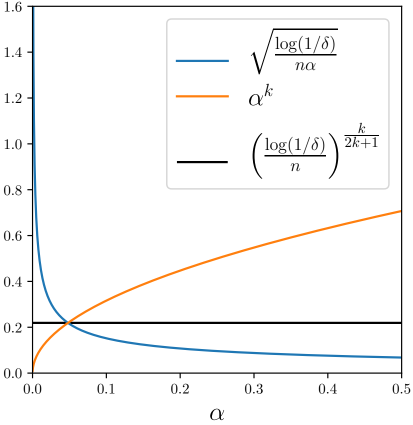
$D(g^\star)$ converges to $\widehat{D}(g^\star)$

- $D(g) = \mathbb{E}[g]^k \, |\mathbb{E}[\ell \mid g] - \mathbb{E}[\ell]|$
  (Weighted loss discrepancy)
- $g^\star = \arg\max_{g \in \mathcal{G}} D(g)$
- $\alpha = \mathbb{E}[g^\star]$
- $D(g^\star) - \widehat{D}(g^\star) \leq \sqrt{\frac{\ln(1/\delta)}{n\alpha}}$
- $D(g^\star) - \widehat{D}(g^\star) \leq \alpha^k$



$D(g^\star)$ converges to $\widehat{D}(g^\star)$

- $D(g) = \mathbb{E}[g]^k \, |\mathbb{E}[\ell \mid g] - \mathbb{E}[\ell]|$
  (Weighted loss discrepancy)
- $g^\star = \arg\max_{g \in \mathcal{G}} D(g)$
- $\alpha = \mathbb{E}[g^\star]$
- $D(g^\star) - \widehat{D}(g^\star) \leq \sqrt{\frac{\ln(1/\delta)}{n\alpha}}$
- $D(g^\star) - \widehat{D}(g^\star) \leq \alpha^k$
- $D(g^\star) - \widehat{D}(g^\star) \leq \max \min_{\alpha} \left(\alpha^k, \sqrt{\frac{\ln(1/\delta)}{n\alpha}} \right)$



$\hat{D}(\hat{g})$ uniformly converges to $D(\hat{g})$

Lemma

Let $\widehat{D}(g) = \hat{\mathbb{E}}[g]^k \left| \hat{\mathbb{E}}[\ell \mid g] - \hat{\mathbb{E}}[\ell] \right|$. *Given* n *data points, let* $\hat{g}$ *be a group with maximum empirical weighted loss discrepancy,* $\hat{g} = \arg\max_{g \in \widehat{\mathcal{G}}} \widehat{D}(g)$.

$\hat{D}(\hat{g})$ uniformly converges to $D(\hat{g})$

Lemma

Let $\widehat{D}(g) = \hat{\mathbb{E}}[g]^k \left| \hat{\mathbb{E}}[\ell \mid g] - \hat{\mathbb{E}}[\ell] \right|$. *Given* n *data points, let* $\hat{g}$ *be a group with maximum empirical weighted loss discrepancy,* $\hat{g} = \arg\max_{g \in \widehat{\mathcal{G}}} \widehat{D}(g)$. *Assume* $\hat{\mathbb{E}}[\ell] = 0$ *and* $\hat{\mathbb{E}}[\ell \mid \hat{g}] \leq \hat{\mathbb{E}}[\ell]$, *then we can represent* $\hat{g}$:

$$\underbrace{\ell_1 \leq \ell_2 \leq \cdots \leq \ell_t}_{\hat{g}} < \ell_{t+1} \leq \cdots \leq \ell_n$$

$\hat{D}(\hat{g})$ uniformly converges to $D(\hat{g})$

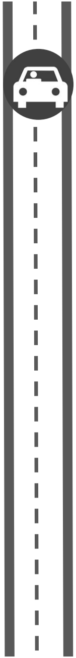
Lemma

Let $\widehat{D}(g) = \hat{\mathbb{E}}[g]^k \left|\hat{\mathbb{E}}[\ell \mid g] - \hat{\mathbb{E}}[\ell]\right|$. *Given n data points, let* $\hat{g}$ *be a group with maximum empirical weighted loss discrepancy,* $\hat{g} = \arg\max_{g \in \widehat{\mathcal{G}}} \widehat{D}(g)$. *Assume* $\hat{\mathbb{E}}[\ell] = 0$ *and* $\hat{\mathbb{E}}[\ell \mid \hat{g}] \leq \hat{\mathbb{E}}[\ell]$*, then we can represent* $\hat{g}$*:*

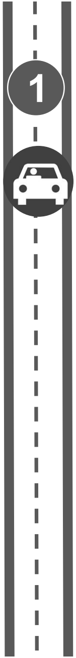
$$\underbrace{\ell_1 \leq \ell_2 \leq \cdots \leq \ell_t}_{\hat{g}} < \ell_{t+1} \leq \cdots \leq \ell_n$$

$$g_u = \{\text{all points with loss less than } u\}$$

$$\sup_{g \in \mathcal{G}} |D(\hat{g}) - \hat{D}(\hat{g})| = \sup_{u \in [0,1]} |D(\hat{g}_u) - \hat{D}(\hat{g}_u)|$$

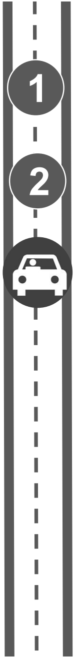


Maximum weighted loss discrepancy, U(w)



Maximum weighted loss discrepancy, U(w)

Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$



Maximum weighted loss discrepancy, U(w)

Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

1. Maximum weighted loss discrepancy, U(w)
2. Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$
3. Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

Loss variance

Hoeffding's inequality

$$\left| \underbrace{\widehat{\mathbb{E}}[\ell]}_{\text{training loss}} - \underbrace{\mathbb{E}[\ell]}_{\text{population loss}} \right| \leq \sqrt{\frac{C_1 \ln(1/\delta)}{n}}$$

Hoeffding's inequality

$$\left| \underbrace{\widehat{\mathbb{E}}[\ell]}_{\text{training loss}} - \underbrace{\mathbb{E}[\ell]}_{\text{population loss}} \right| \leq \sqrt{\frac{C_1 \ln(1/\delta)}{n}}$$

Bennett's inequality

$$\left| \underbrace{\widehat{\mathbb{E}}[\ell]}_{\text{training loss}} - \underbrace{\mathbb{E}[\ell]}_{\text{population loss}} \right| \leq \sqrt{\frac{C_1 \ln(1/\delta) \mathrm{Var}[\ell]}{n}} + \frac{C_2 \ln(1/\delta)}{n}$$

Bennett's inequality

$$\Bigg|\underbrace{\widehat{\mathbb{E}}[\ell]}_{\text{training loss}} - \underbrace{\mathbb{E}[\ell]}_{\text{population loss}}\Bigg| \leq \sqrt{\frac{C_1 \mathrm{Var}[\ell]}{n}} + \frac{C_2}{n}$$

[Bennett, 1962]
[Maurer and Pontil, 2009]
[Mnih et al., 2008]
[Audibert et al., 2009]
[Shivaswamy and Jebara, 2010]
[Namkoong and Duchi, 2017]

Bennett's inequality

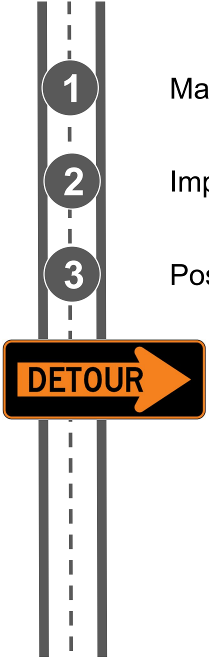
$$\left| \underbrace{\widehat{\mathbb{E}}[\ell]}_{\text{training loss}} - \underbrace{\mathbb{E}[\ell]}_{\text{population loss}} \right| \leq \sqrt{\frac{C_1 \mathrm{Var}[\ell]}{n}} + \frac{C_2}{n}$$

[Bennett, 1962]
[Maurer and Pontil, 2009] $\Longrightarrow$ Empirical Bernstein Bounds and Sample Variance Penalization
[Mnih et al., 2008]
[Audibert et al., 2009]
[Shivaswamy and Jebara, 2010]
[Namkoong and Duchi, 2017]

Bennett's inequality

$$\Bigg| \underbrace{\widehat{\mathbb{E}}[\ell]}_{\text{training loss}} - \underbrace{\mathbb{E}[\ell]}_{\text{population loss}} \Bigg| \leq \sqrt{\frac{C_1 \mathrm{Var}[\ell]}{n}} + \frac{C_2}{n}$$

[Bennett, 1962]
[Maurer and Pontil, 2009] $\implies$ Empirical Bernstein Bounds and Sample Variance Penalization
[Mnih et al., 2008]
[Audibert et al., 2009]
[Shivaswamy and Jebara, 2010]
[Namkoong and Duchi, 2017] $\implies$ Variance-based regularization with convex objectives



1. Maximum weighted loss discrepancy, U(w)
2. Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$
3. Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

Loss variance → Test error

Average individual discrepancy

$$\begin{aligned}\mathsf{Var}[\ell] &= \mathbb{E}\left[\left(\ell(z) - \mathbb{E}[\ell]\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}_{z,z' \sim p^\star}\left[\left(\ell(z) - \ell(z')\right)^2\right]\end{aligned}$$

1 Maximum weighted loss discrepancy, U(w)

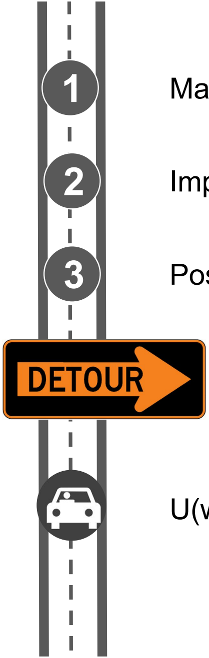
2 Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

3 Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

DETOUR

Loss variance

- Test error
- Average individual discrepancy



Maximum weighted loss discrepancy, U(w)

Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

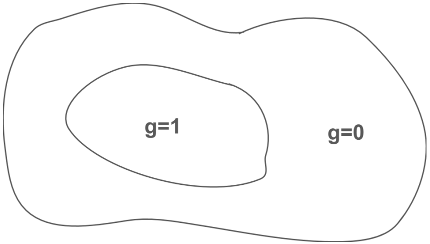
Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

Loss variance → Test error

Loss variance → Average individual discrepancy

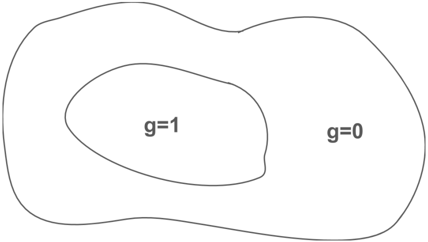
U(w) for $w(g) = \mathbb{E}[g]^{\frac{1}{2}}$ $\longleftrightarrow$ Loss variance

law of total variance



$$\mathrm{Var}[\mathbb{E}[\ell \mid g]] \leq \mathrm{Var}[\ell]$$

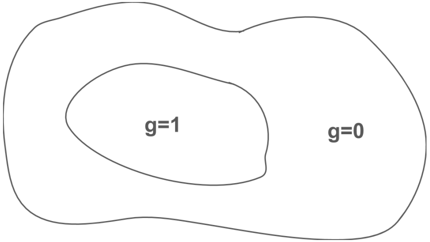
law of total variance



$$\mathrm{Var}[\mathbb{E}[\ell \mid g]] \leq \mathrm{Var}[\ell]$$

$$\mathbb{E}[g](\mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell])^2 + \mathbb{E}[1 - g](\mathbb{E}[\ell \mid g = 0] - \mathbb{E}[\ell])^2 \leq \mathrm{Var}[\ell]$$

law of total variance

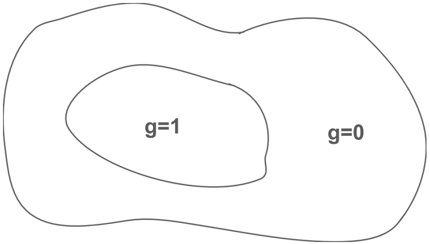


$$\mathsf{Var}[\mathbb{E}[\ell \mid g]] \leq \mathsf{Var}[\ell]$$

$$\mathbb{E}[g](\mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell])^2 + \mathbb{E}[1 - g](\mathbb{E}[\ell \mid g = 0] - \mathbb{E}[\ell])^2 \leq \mathsf{Var}[\ell]$$

$$\underbrace{\sqrt{\mathbb{E}[g]}}_{\text{weighted}} \underbrace{|\mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell]|}_{\text{loss discrepancy}} \leq \sqrt{\mathsf{Var}[\ell]}$$

law of total variance



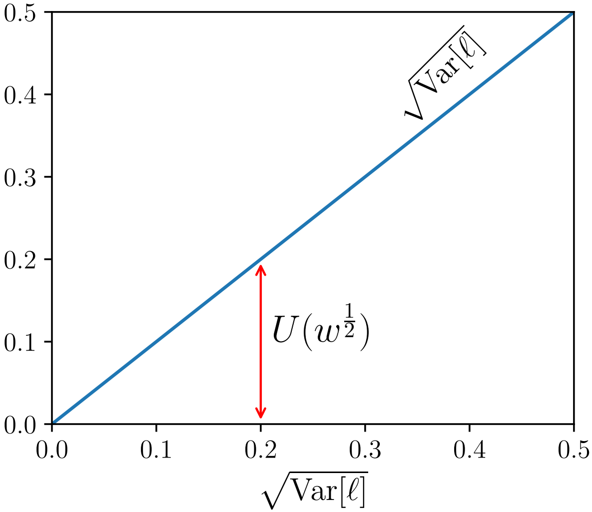
$$\mathsf{Var}[\mathbb{E}[\ell \mid g]] \leq \mathsf{Var}[\ell]$$

$$\mathbb{E}[g](\mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell])^2 + \mathbb{E}[1 - g](\mathbb{E}[\ell \mid g = 0] - \mathbb{E}[\ell])^2 \leq \mathsf{Var}[\ell]$$

$$\underbrace{\sqrt{\mathbb{E}[g]}}_{\text{weighted}} \underbrace{|\mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell]|}_{\text{loss discrepancy}} \leq \sqrt{\mathsf{Var}[\ell]}$$

Proposition

Let $w^{\frac{1}{2}}(g) = \mathbb{E}[g]^{\frac{1}{2}}$ *then* $U(w^{\frac{1}{2}}) \leq \sqrt{Var[\ell]}$.



0.0
0.1
0.2
0.3
0.4
0.5
$\sqrt{\mathrm{Var}[\ell]}$
$U(w^{\frac{1}{2}})$
$\sqrt{\mathrm{Var}[\ell]}$

Proposition

Let $w^{\frac{1}{2}}(g) = \mathbb{E}[g]^{\frac{1}{2}}$ *then* $\sqrt{Var[\ell]} \leq U(w^{\frac{1}{2}})\sqrt{2 - 4\ln(U(w^{\frac{1}{2}}))}$

Proposition

Let $w^{\frac{1}{2}}(g) = \mathbb{E}[g]^{\frac{1}{2}}$ *then* $\sqrt{Var[\ell]} \leq U(w^{\frac{1}{2}})\sqrt{2 - 4\ln(U(w^{\frac{1}{2}}))}$

Proof idea.

- $\mathbb{P}[\ell \geq u] \leq \frac{U(w^{\frac{1}{2}})}{u^2}$

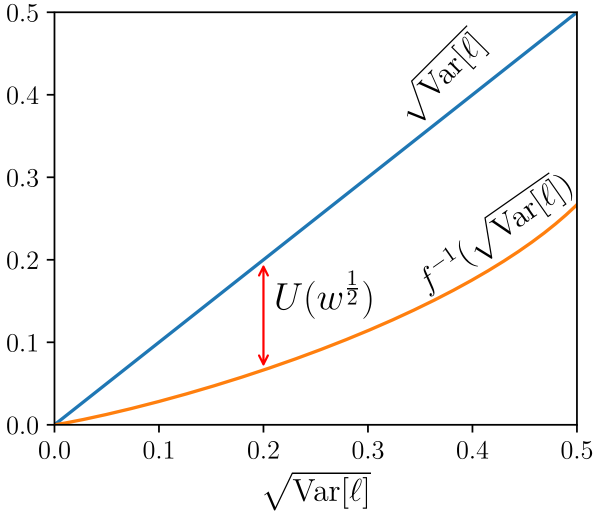
Proposition

Let $w^{\frac{1}{2}}(g) = \mathbb{E}[g]^{\frac{1}{2}}$ *then* $\sqrt{Var[\ell]} \leq U(w^{\frac{1}{2}})\sqrt{2 - 4\ln(U(w^{\frac{1}{2}}))}$

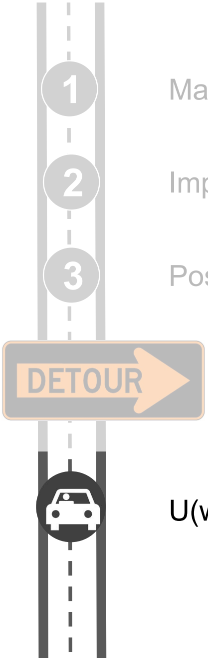
Proof idea.

- $\mathbb{P}[\ell \geq u] \leq \frac{U(w^{\frac{1}{2}})}{u^2}$
- $\mathrm{Var}[\ell] = \int u\mathbb{P}[\ell \geq u]$

$\square$



$f(x) = x\sqrt{2 - 4\ln(x)}$



Maximum weighted loss discrepancy, U(w)

Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

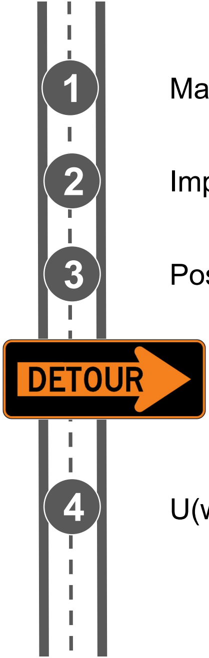
Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

Loss variance

Test error

Average individual discrepancy

U(w) for $w(g) = \mathbb{E}[g]^{\frac{1}{2}}$ $\longleftrightarrow$ Loss variance



1. Maximum weighted loss discrepancy, U(w)
2. Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$
3. Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

Loss variance → Test error

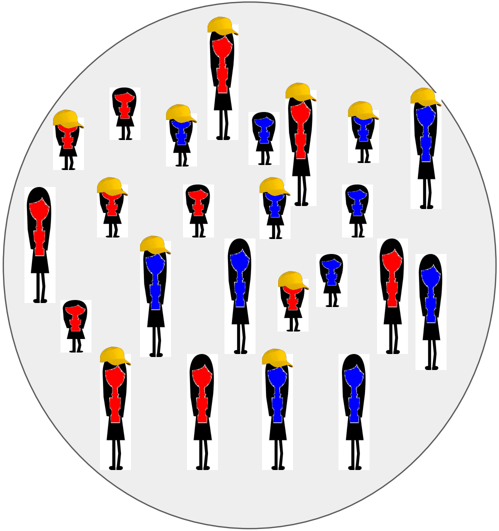
Loss variance → Average individual discrepancy

4. U(w) for $w(g) = \mathbb{E}[g]^{\frac{1}{2}}$ ⟷ Loss variance

No prior information

Handling prior information

A is given
A = [height, color]

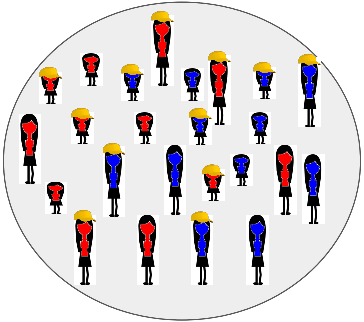


$$U(w) = \max_{g \in \mathcal{G}} w(g) \left| \mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell] \right|$$

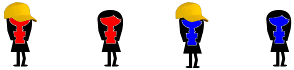
$$U(w) = \max_{g \in \mathcal{G}} w(g) \Big| \mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell] \Big|$$

$$w(g) = \begin{cases} > 0 & g \in \mathcal{G}_A \\ 0 & o.w. \end{cases}$$

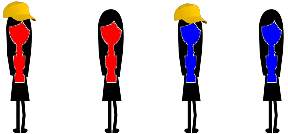
All possible groups on A = [height, color]



Coarse loss variance



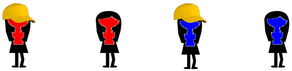
$\ell = 1$ $\ell = 0$ $\ell = 1$ $\ell = 0$



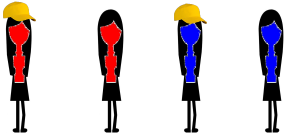
$\ell = 1$ $\ell = 0$ $\ell = 1$ $\ell = 0$

$$\mathrm{Var}[\ell] = 0.25$$

Coarse loss variance



$\ell = 1$ $\ell = 0$ $\ell = 1$ $\ell = 0$

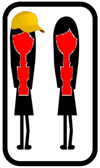


$\ell = 1$ $\ell = 0$ $\ell = 1$ $\ell = 0$

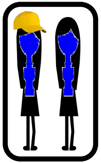
$$\mathrm{Var}[\ell] = 0.25$$

$\mathbb{E}[\ell \mid A = [\begin{matrix} short, \\ red \end{matrix}]] = 0.5$

$\mathbb{E}[\ell \mid A = [\begin{matrix} short, \\ blue \end{matrix}]] = 0.5$



$\mathbb{E}[\ell \mid A = [\begin{matrix} tall, \\ red \end{matrix}]] = 0.5$



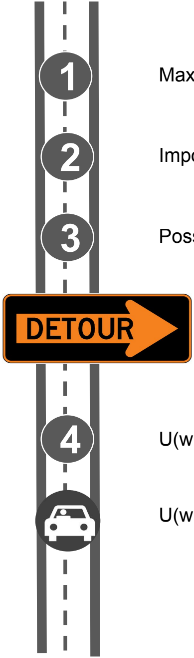
$\mathbb{E}[\ell \mid A = [\begin{matrix} tall, \\ blue \end{matrix}]] = 0.5$

$$\mathrm{Var}[\mathbb{E}[\ell \mid A]] = 0$$

Coarse loss variance

$$U(w) \text{ for } w(g) = \begin{cases} \mathbb{E}[g]^{\frac{1}{2}} & g \in \mathcal{G}_A \\ 0 & o.w. \end{cases} \iff \mathrm{Var}[\mathbb{E}[\ell \mid A]]$$

All possible groups on A = [height, color]



Maximum weighted loss discrepancy, U(w)

Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

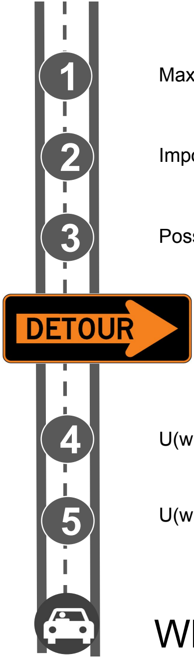
Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

Loss variance → Test error

Loss variance → Average individual discrepancy

U(w) for $w(g) = \mathbb{E}[g]^{\frac{1}{2}}$ ↔ Loss variance

U(w) for $w(g) = \begin{cases} \mathbb{E}[g]^{\frac{1}{2}} & g \in \mathcal{G}_A \\ 0 & o.w. \end{cases}$ ↔ Coarse loss variance



1. Maximum weighted loss discrepancy, U(w)
2. Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$
3. Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

Loss variance

Test error

Average individual discrepancy

4. U(w) for $w(g) = \mathbb{E}[g]^{\frac{1}{2}}$ $\longleftrightarrow$ Loss variance
5. U(w) for $w(g) = \begin{cases} \mathbb{E}[g]^{\frac{1}{2}} & g \in \mathcal{G}_A \\ 0 & o.w. \end{cases}$ $\longleftrightarrow$ Coarse loss variance

What should be the weighting function?

Questions?

1. Maximum weighted loss discrepancy, U(w)

2. Impossible to audit U(w) for $w(g) = \mathbb{I}[\mathbb{E}[g] > 0]$

3. Possible to audit U(w) for $w(g) = \mathbb{E}[g]^k$

DETOUR

Loss variance → Test error

Loss variance → Average individual discrepancy

4. U(w) for $w(g) = \mathbb{E}[g]^{\frac{1}{2}}$ ⟷ Loss variance

5. U(w) for $w(g) = \begin{cases} \mathbb{E}[g]^{\frac{1}{2}} & g \in \mathcal{G}_A \\ 0 & o.w. \end{cases}$ ⟷ Coarse loss variance

6. What should be the weighting function?

References

J. Audibert, R. Munos, and C. Szepesv'ari. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association (JASA)*, 57(297):33–45, 1962.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

V. Mnih, C. Szepesv'ari, and J. Audibert. Empirical berstein stopping. In *International Conference on Machine Learning (ICML)*, 2008.

H. Namkoong and J. Duchi. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

P. Shivaswamy and T. Jebara. Empirical Bernstein boosting. In *Artificial Intelligence and Statistics (AISTATS)*, pages 733–740, 2010.