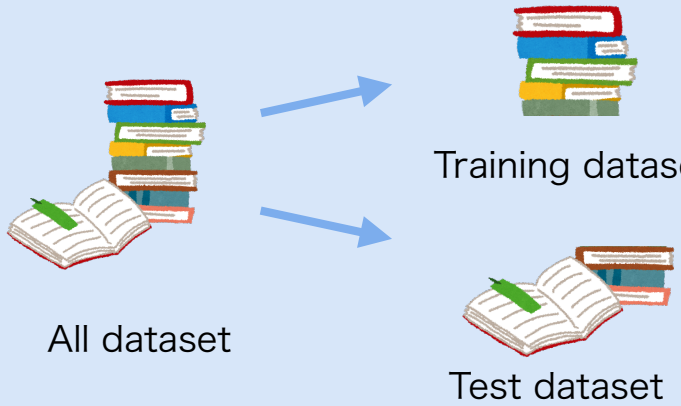
Estimating the Generalization Gap in Deep Neural Networks using the Network Information Criterion

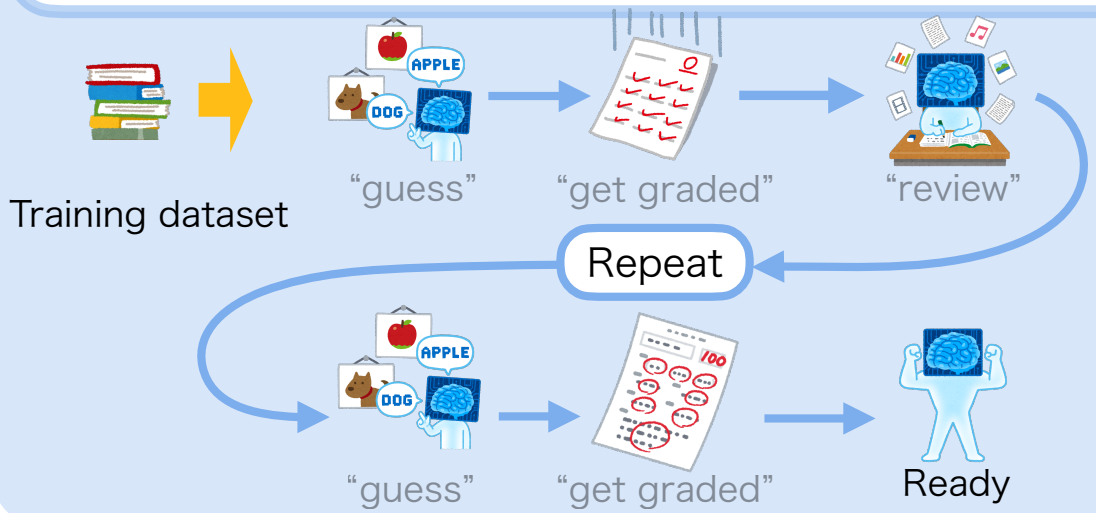
Sutashu TOMONAGA

Training and Evaluation of a Model

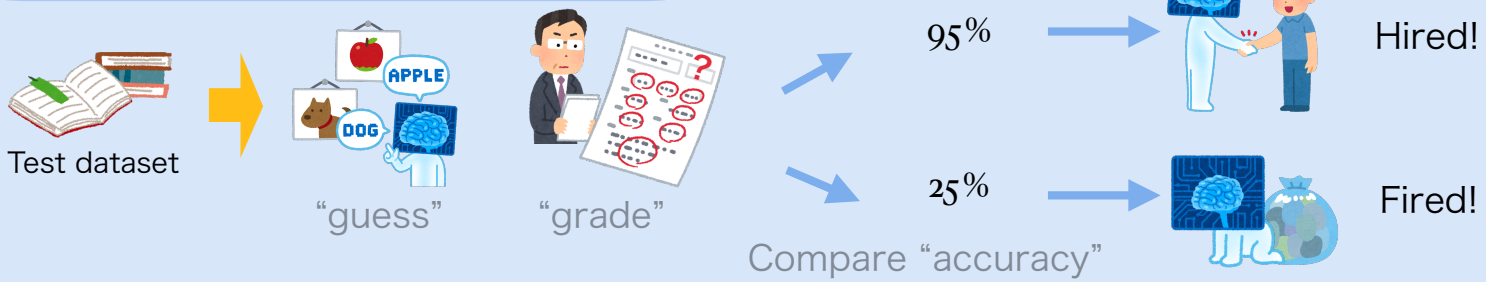
1 Split given data for Training and Testing



2 Model repeats tasks: “guess answers”→ “get graded” → “review answers”



3 Measure the the model’s error using the Test dataset

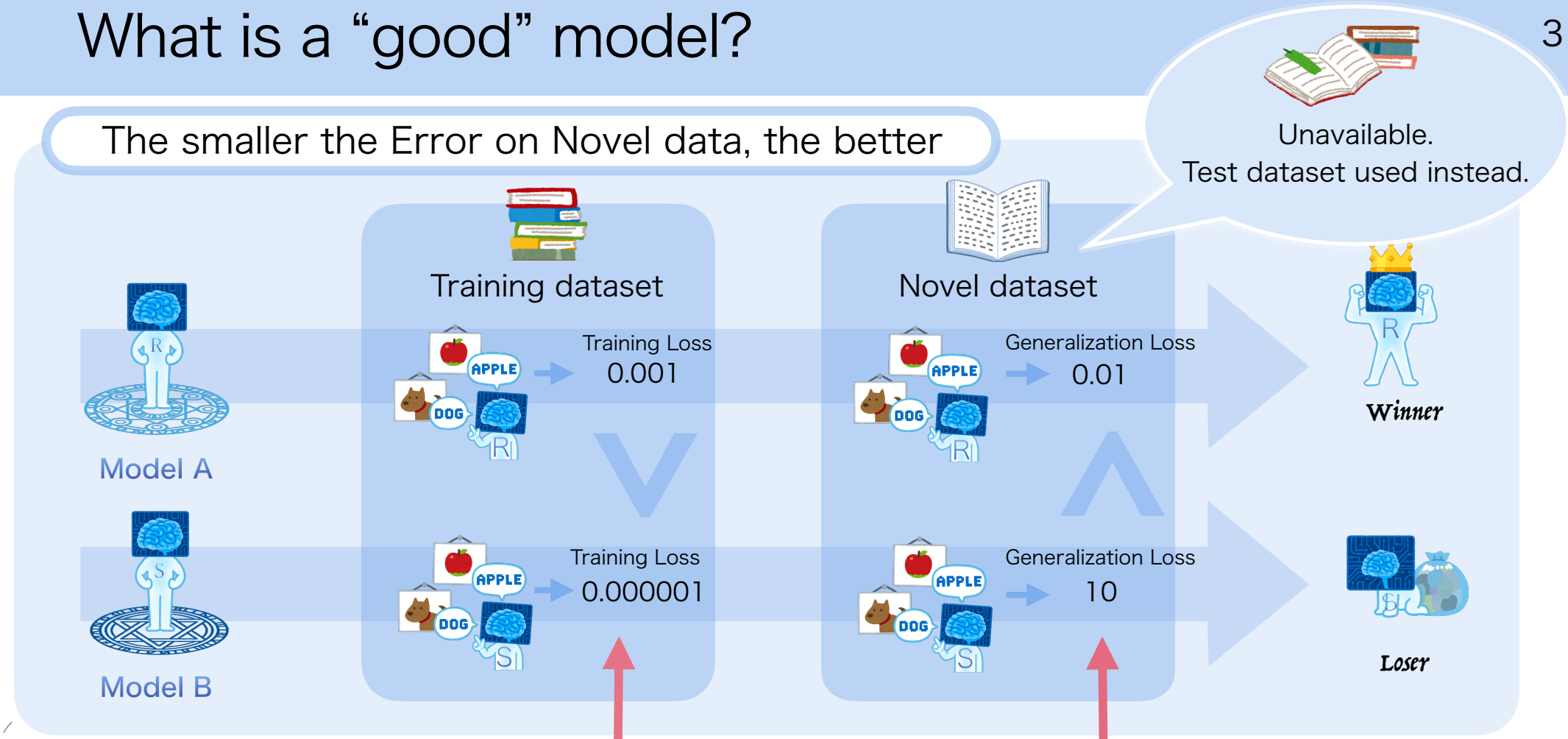


note

To “guess answer” :
to obtain output of model
To “grade” :
to measure the error of model
To “review” :
to modify parameters

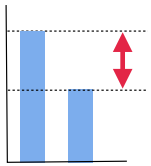
What is a “good” model?

The smaller the Error on Novel data, the better



Models are evaluated with Test dataset
(dataset not used for training)

Gap = Generalization Gap



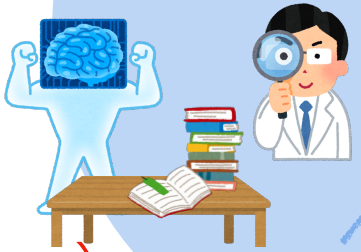
Ideal Model Selection

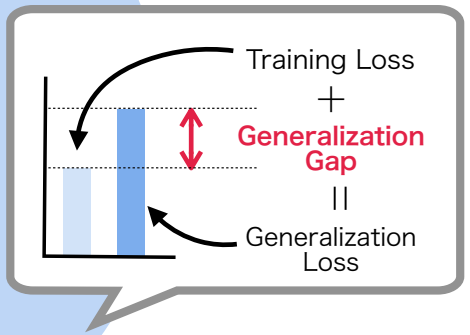
IDEAL : able to use more (or all) of the given data for training…

But if we do, we can't obtain the Generalization Loss…

If we can estimate the **Generalization Gap** using ONLY the Training dataset… then we know the Generalization Loss

→ **NIC**





NIC (Network Information Criterion)

Estimator of the Generalization Gap

$$NIC = \frac{1}{N}\mathrm{Tr}(H^{-1}C)$$

N : Number of samples

ℓ : Loss function

$$H(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) = \mathbb{E}_q\left[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})\right]$$ Expected Hessian

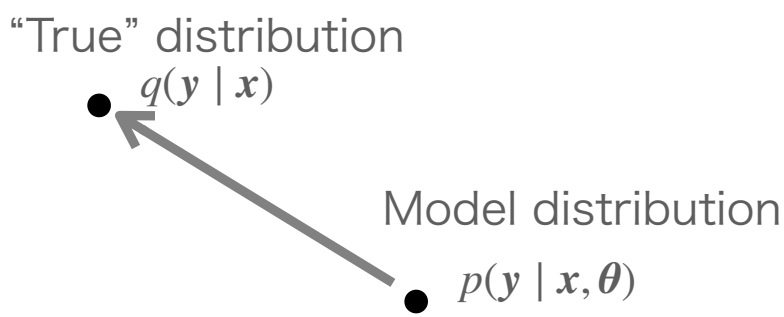
$$C(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) = \mathbb{E}_q\left[\left(\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})\right)\left(\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})\right)^{\mathrm{T}}\right]$$

Expected gradient covariance

What does Machine Learning Optimize?

Goal: Approximate “true” distribution with a parametric model

- “True” distribution: Stochastic System (Probability Distribution $q(\boldsymbol{y} \mid \boldsymbol{x})$)
- Model distribution: Statistical Model (Probability Distribution $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$)



- Minimize Expected Loss $D(q, p(\boldsymbol{\theta})) = \mathbb{E}_q \left[\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \right]$

$$\mathbb{E}_q \left[\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \right] = \int \ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) q(\boldsymbol{x}) q(\boldsymbol{y} \mid \boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y} = \int \ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) q(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}$$

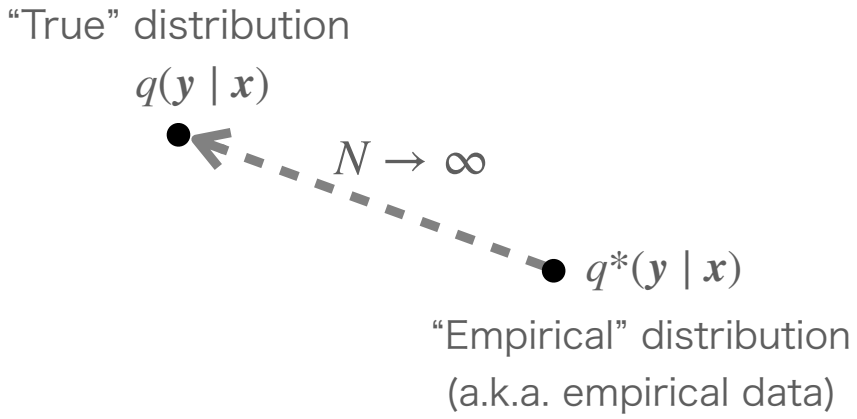
$$\boldsymbol{\theta}_{opt} \in \arg\min_{\boldsymbol{\theta}} D(q, p(\boldsymbol{\theta}))$$

Empirical Distribution

- "True" Distribution $q(\boldsymbol{y} \mid \boldsymbol{x})$: Unknown
- "Empirical" Distribution $q^*(\boldsymbol{y} \mid \boldsymbol{x})$: Known

$$q^*(\boldsymbol{x}, \boldsymbol{y}) \equiv \frac{1}{N} \sum_{i=1}^{N} \delta\left(\boldsymbol{x} - \boldsymbol{x}_i, \boldsymbol{y} - \boldsymbol{y}_i\right)$$

where $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are samples obtained from $q(\boldsymbol{y} \mid \boldsymbol{x})$



- When $N \to \infty$,

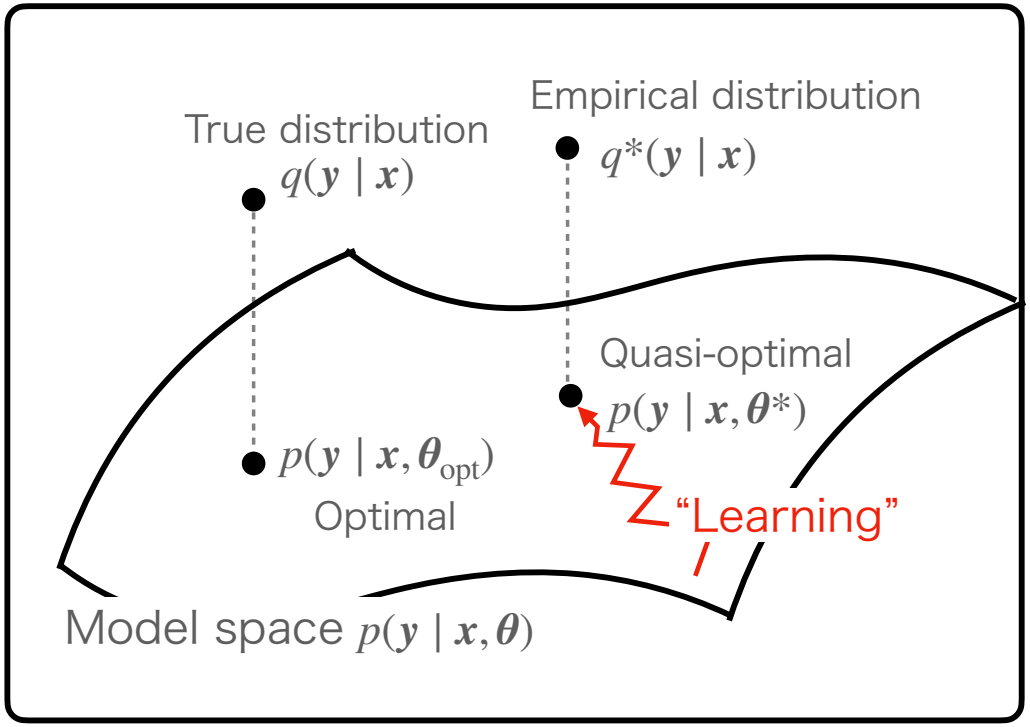
$$q^*(\boldsymbol{y} \mid \boldsymbol{x}) \to q(\boldsymbol{y} \mid \boldsymbol{x})$$

Optimizing to the Empirical Distribution

- Only the Empirical distribution is available
  - Optimum $\boldsymbol{\theta}_{opt}$ is unobtainable
  - Search for Quasi-optimum $\boldsymbol{\theta}^*$

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}} D(q^*, p(\boldsymbol{\theta}))$$

$$q^*(\boldsymbol{x}, \boldsymbol{y}) \equiv \frac{1}{N} \sum_{i=1}^{N} \delta\left(\boldsymbol{x} - \boldsymbol{x}_i, \boldsymbol{y} - \boldsymbol{y}_i\right)$$



Model Selecting

- Select model with best performance to Novel data
  (Performance measured by “Generalization Loss” $D(q, p(\boldsymbol{\theta}*))$: unknown)

- In practice, split $q*$ into q^*_{train} and q^*_{test} (randomly)

  - Train model with q^*_{train} ($\boldsymbol{\theta}* \in \arg\min_{\boldsymbol{\theta}} D(q^*_{train}, p(\boldsymbol{\theta}))$)

  - Use $D(q^*_{test}, p(\boldsymbol{\theta}*))$ as an estimator of Generalization Loss

    … Only because “it works”

    … Wasting data for testing

Estimating Generalization Loss in a different way

- We want:

  Generalization Loss $D(q, p(\boldsymbol{\theta}^*))$

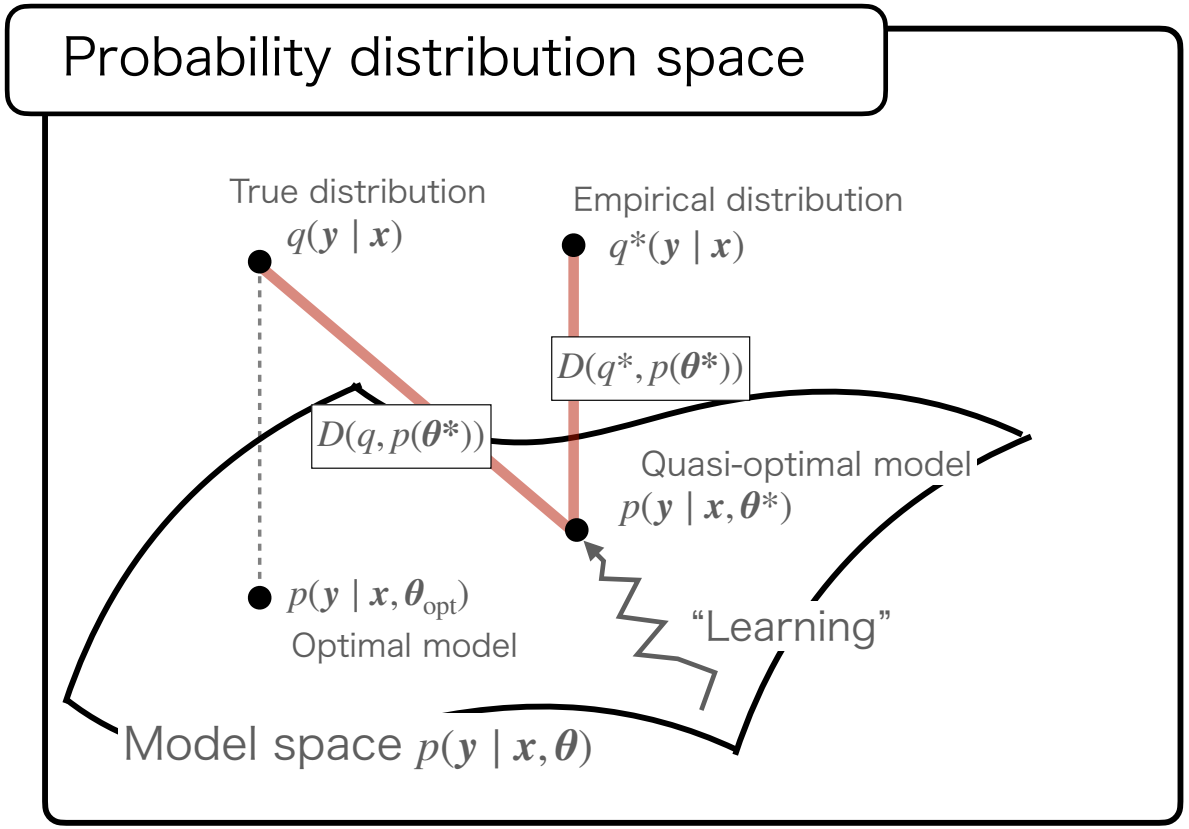
- We have:

  Train Loss $D(q^*, p(\boldsymbol{\theta}^*))$

- Can we estimate:

  Generalization Gap $\mathscr{G}$?

  $D(q, p(\boldsymbol{\theta}^*)) - D(q^*, p(\boldsymbol{\theta}^*))$

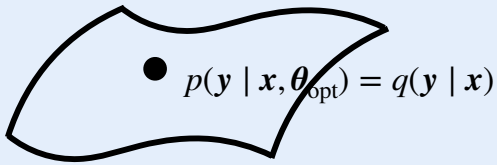


➡Yes, (under certain conditions, using Information Criterion)

Information Criterion

Akaike's Information Criterion

- If there exists θ_{opt} s.t. $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{opt}}) = q(\boldsymbol{y} \mid \boldsymbol{x})$



$$D(q, p(\boldsymbol{\theta}^*)) = AIC(p) + D(q^*, p(\boldsymbol{\theta}^*))$$

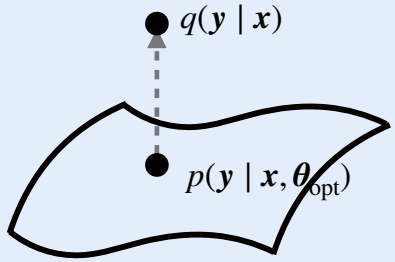
$$AIC(p) = \frac{d}{N}$$

N : Number of samples

d : degrees of freedom (number of parameters) of p

Network Information Criterion

- If not,



$$D(q, p(\boldsymbol{\theta}^*)) = NIC\left(p(\boldsymbol{\theta}^*)\right) + D(q^*, p(\boldsymbol{\theta}^*))$$

$$NIC(p) = \frac{1}{N}\mathrm{Tr}(H^{-1}C)$$

$$H\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right) = \mathbb{E}_q\left[\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right)\right]$$

$$C\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right) = \mathbb{E}_q\left[\left(\nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right)\right)\left(\nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right]$$

What is H and C ?

$$NIC\left(p\left(\boldsymbol{\theta}\right)\right) = \frac{1}{N}\mathrm{Tr}\left(H^{-1}\left(\boldsymbol{\theta}\right)C\left(\boldsymbol{\theta}\right)\right)$$

Expected Hessian matrix of Loss function

$$H\left(\boldsymbol{\theta}\right) = \mathbb{E}_q\left[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right]$$

Expected Gradient Covariance matrix of Loss function

$$C\left(\boldsymbol{\theta}\right) = \mathbb{E}_q\left[\left(\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right)\left(\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right]$$

In practice,

$$NIC^*\left(p\left(\boldsymbol{\theta}\right)\right) = \frac{1}{N}\mathrm{Tr}\left(H^{*-1}\left(\boldsymbol{\theta}\right)C^*\left(\boldsymbol{\theta}\right)\right)$$

Empirical…

$$H^*\left(\boldsymbol{\theta}\right) = \mathbb{E}_{q^*}\left[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right]$$

Empirical…

$$C^*\left(\boldsymbol{\theta}\right) = \mathbb{E}_{q^*}\left[\left(\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right)\left(\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right]$$

Expectation (weighted integral)

$$\mathbb{E}_q\left[\,f\,\right] = \int f\,q(\boldsymbol{x})\;q(\boldsymbol{y}\mid\boldsymbol{x})d\boldsymbol{x}d\boldsymbol{y}$$

$$= \int f\,q(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}$$

Using empirical distribution…

$$\mathbb{E}_{q^*}\left[\,f\,\right] = \int f\,q^*(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}$$

$$q^*(\boldsymbol{x},\boldsymbol{y}) \equiv \frac{1}{N}\sum_{i=1}^{N}\delta\left(\boldsymbol{x}-\boldsymbol{x}_i,\boldsymbol{y}-\boldsymbol{y}_i\right)$$

Other Information matrices

Fisher Information Matrix

$$
\begin{aligned}
F(\boldsymbol{\theta}) &= \mathbb{E}_p\left[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right] \\
&= \mathbb{E}_p\left[\left(\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right)\left(\nabla_{\boldsymbol{\theta}}\,\ell\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right]
\end{aligned}
$$

Expectation (weighted integral)

$$
\mathbb{E}_p\left[\,f\,\right] = \int f\; q(\boldsymbol{x})\; p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}d\boldsymbol{y}
$$

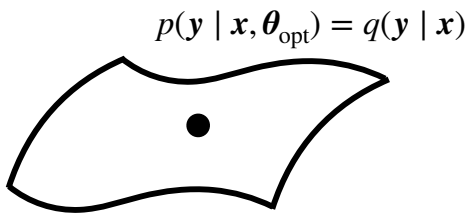
Empirical NIC converges to AIC

- If $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{opt}}) = q(\boldsymbol{y} \mid \boldsymbol{x})$, then $F = H = C$
- If $N \to \infty$ then, $H = H^*, C = C^*$

$$H(\boldsymbol{\theta}) = \mathbb{E}_q\left[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})\right]$$

$$C(\boldsymbol{\theta}) = \mathbb{E}_q\left[\left(\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})\right)\left(\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})\right)^{\mathrm{T}}\right]$$

$$\begin{aligned} F(\boldsymbol{\theta}) &= \mathbb{E}_p\left[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})\right] \\ &= \mathbb{E}_p\left[\left(\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})\right)\left(\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})\right)^{\mathrm{T}}\right] \end{aligned}$$



When both conditions are satisfied,

$$NIC^* = \frac{1}{N}\mathrm{Tr}(H^{*-1}C^*) = \frac{1}{N}\mathrm{Tr}(H^{-1}C) = \frac{1}{N}\mathrm{Tr}(F^{-1}F) = \frac{1}{N}d = AIC$$

- $NIC^* = AIC$ when conditions are met, or, NIC^* is deviation of AIC when not ideal
- Information about generalization capability of a model is contained in the diagonal values
  - Non-diagonal elements emerge to compensate for the deviation

NIC is a generalization of AIC

Known, used for empirical NIC

$H^*(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^*)$ $C^*(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^*)$

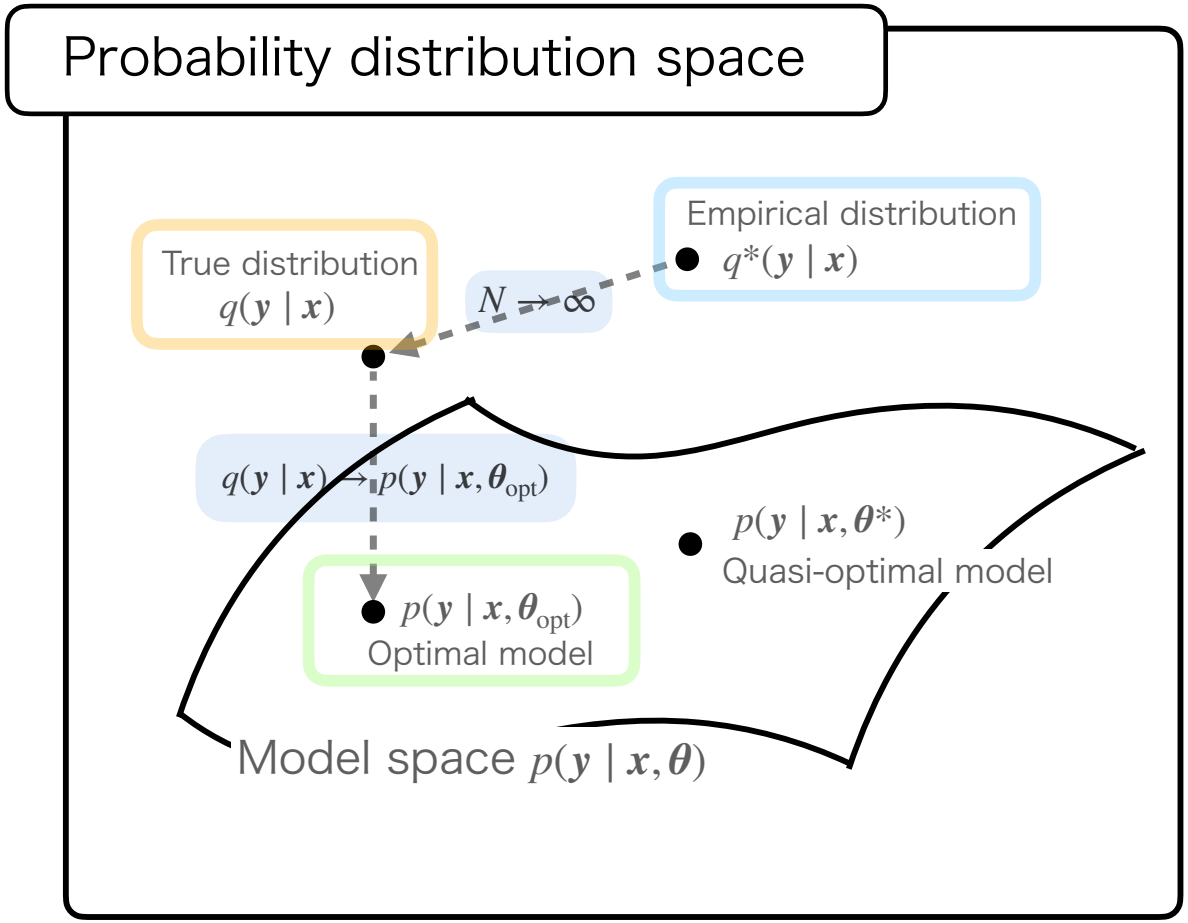
If $N \to \infty$, then $H = H^*, C = C^*$

Unknown

$H(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}_{\mathrm{opt}})$ $C(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}_{\mathrm{opt}})$

If $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{opt}}) = q(\boldsymbol{y} \mid \boldsymbol{x})$, then $F = H = C$

Unknown $F(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}_{\mathrm{opt}})$

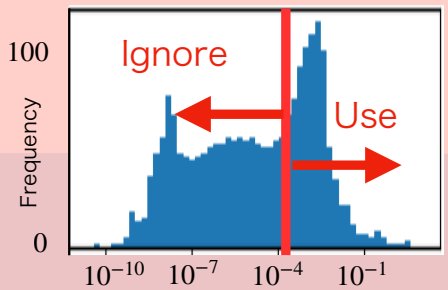


Problem with NIC in Deep Neural Networks

- Deep Neural Networks have MANY variables (thousands ~ billions)
  - Allows great approximation capabilities
- However, H^* of DNN quickly degenerates
  - Unable to compute $NIC^* = \frac{1}{N}\mathrm{Tr}(H^{*-1}C^*)$

Past Research

Can NIC^* estimate the Generalization Gap if degenerate Eigenspaces are ignored?



Histogram of Absolute Eigenvalues of H^*

How degenerate Eigenspaces are ignored

- Diagonalize H

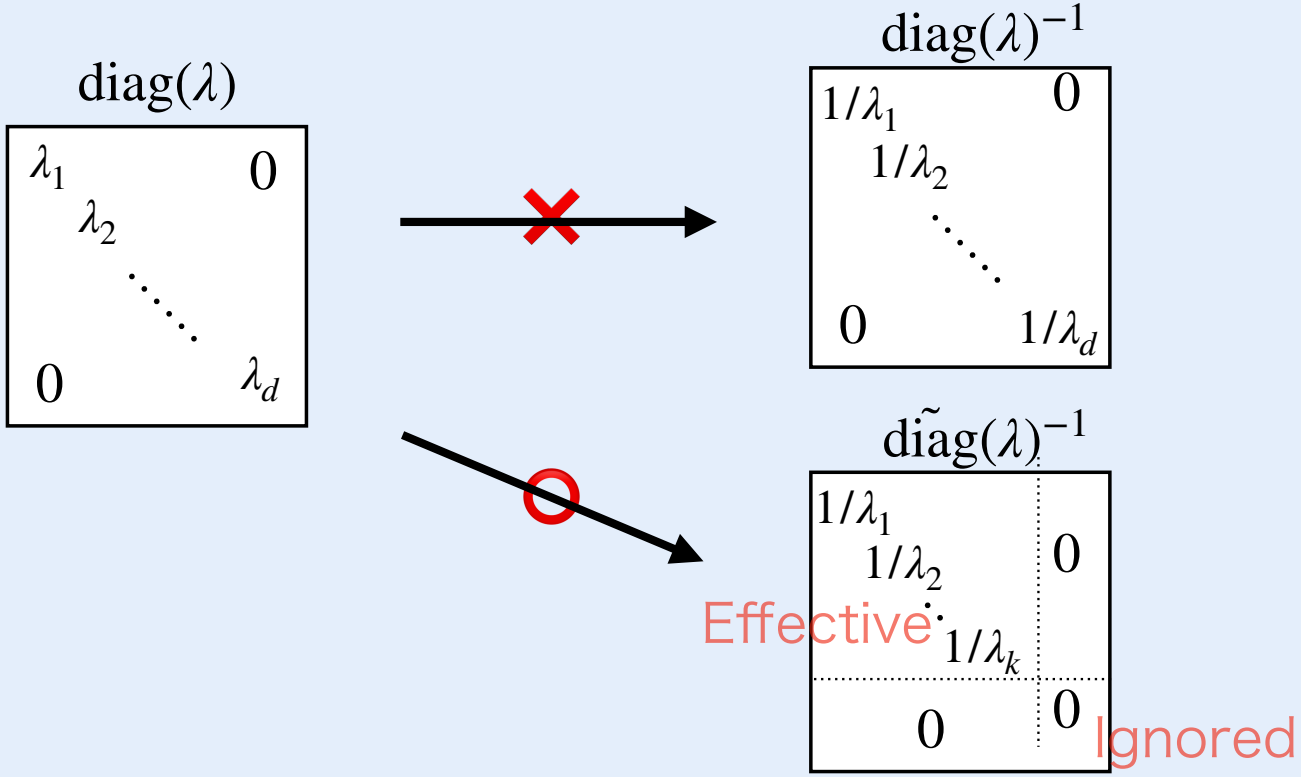
$$H = RAR^{-1} = R\ \mathrm{diag}(\lambda)\ R^{-1}$$

- Inverse of H can be denoted as

$$\begin{aligned} H^{-1} &= (R\ \mathrm{diag}(\lambda)\ R^{-1})^{-1} \\ &= (R^{-1})^{-1}\ \mathrm{diag}(\lambda)^{-1}\ R^{-1} \\ &= R\ \mathrm{diag}(\lambda^{-1})\ R^{-1} \end{aligned}$$

How degenerate Eigenspaces are ignored

- When computing $\mathrm{diag}(\lambda)^{-1}$, only keep inverse values larger than threshold

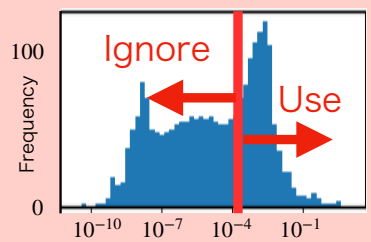


Past research: Results and conclusions

Past Research

Can NIC^* estimate the Generalization Gap if degenerate Eigenspaces are ignored?

- Result: NIC^* works for DNNs when using $NIC^*\left(q^*_{test}, p\left(\boldsymbol{\theta}\right)\right)$, but not $NIC^*\left(q^*_{train}, p\left(\boldsymbol{\theta}\right)\right)$
  - We still have to split q^* into q^*_{train} and q^*_{test}
  - Why $NIC^*\left(q^*_{train}, p\left(\boldsymbol{\theta}\right)\right)$ doesn't work is a mystery
  - Threshold is arbitrary



Histogram of Absolute Eigenvalues of H^*

Question

- How about using eigenvalues of C^* instead of H^* , to determine effective/ignored Eigenvalues?

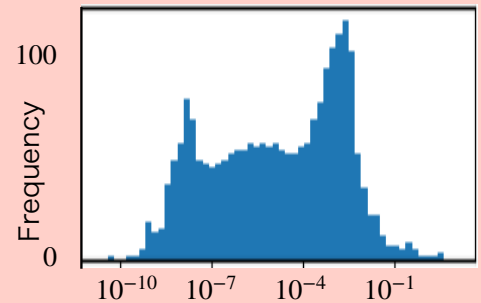
Empirical Expected Gradient Covariance matrix of Loss function

$$C^*(\boldsymbol{\theta}) = \mathbb{E}_{q^*}\left[\left(\nabla_{\boldsymbol{\theta}}\, \ell\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right)\right)\left(\nabla_{\boldsymbol{\theta}}\, \ell\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right]$$

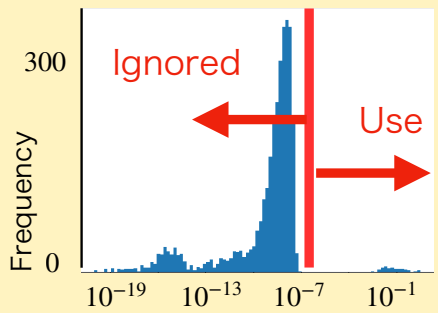
$$q^*(\boldsymbol{x}, \boldsymbol{y}) \equiv \frac{1}{N}\sum_{i=1}^{N} \delta\left(\boldsymbol{x} - \boldsymbol{x}_i, \boldsymbol{y} - \boldsymbol{y}_i\right) \qquad \mathbb{E}_{q^*}\left[\, f \,\right] = \int f\, q^*(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}$$

- Rank of C^* depends on number of samples
  - C^* has a separated distribution

… What does “use eigenvalues of C^*” mean?



Histogram of Absolute Eigenvalues of H^*



Histogram of Absolute Eigenvalues of C^*

Transforming H with Eigenbases of C

$$\mathrm{tr}\left(H^{-1}C\right) = \mathrm{tr}\left(H^{-1}\left(R_C D_C R_C^{-1}\right)\right)$$

Eliminate Degenerate Space $\rightarrow$ $\mathrm{tr}\left(H^{-1}\left(R\tilde{I}_C D_C \tilde{I}_C R^{-1}\right)\right)$

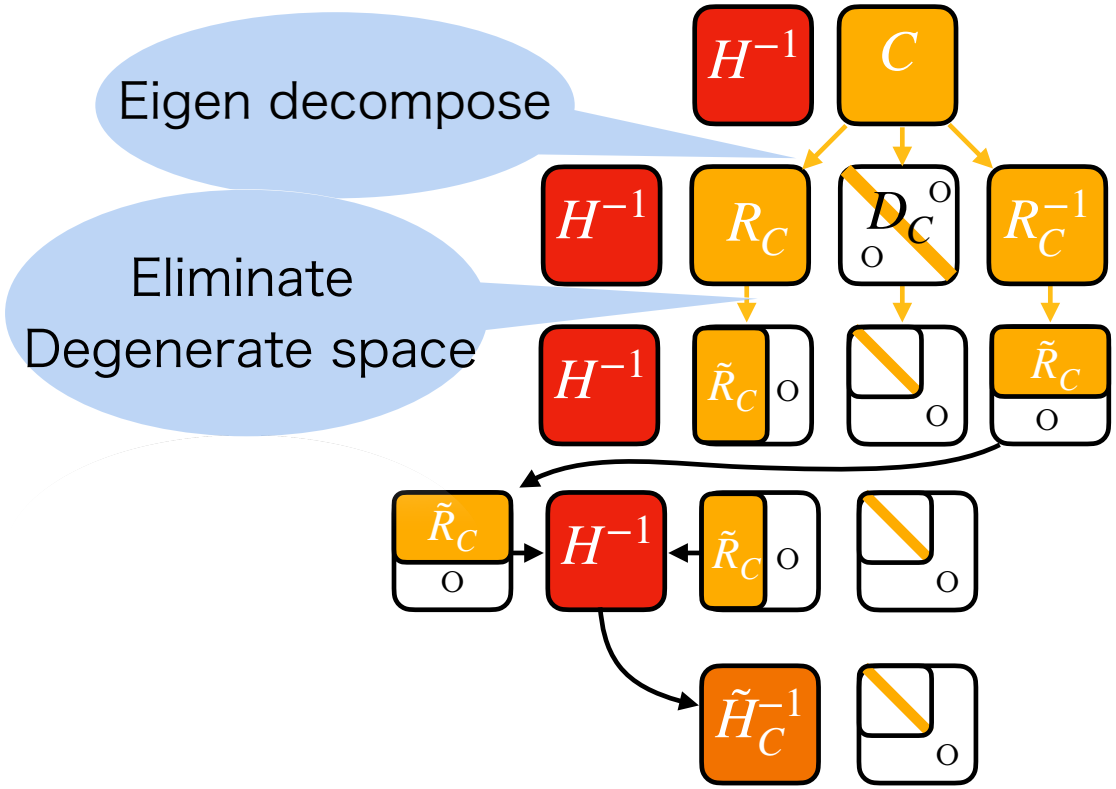
$$= \mathrm{tr}\left(H^{-1}\left(R\tilde{I}_C\right) D_C \left(\tilde{I}_C R^{-1}\right)\right)$$

$$= \mathrm{tr}\left(H^{-1}\tilde{R}_C D_C \tilde{R}_C^{-1}\right)$$

$$= \mathrm{tr}\left(\tilde{R}_C^{-1} H^{-1} \tilde{R}_C D_C\right)$$

$$= \mathrm{tr}\left(\left(\tilde{R}_C H \tilde{R}_C^{-1}\right)^{-1} D_C\right)$$

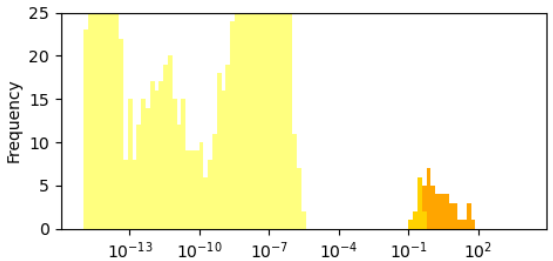
$$= \mathrm{tr}\left(\tilde{H}_C^{-1} D_C\right)$$



Transforming H with Eigenbases of C

Effective space of C^*

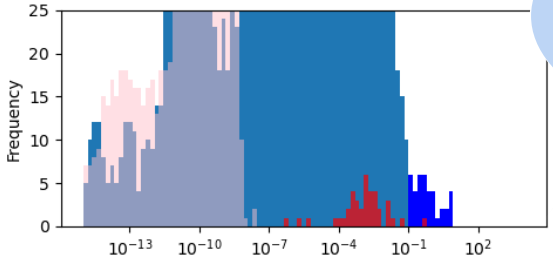
Eliminated space of C^*



Histogram of Absolute Eigenvalues of C^*

Effective space of $\tilde{H}^*_{C^*}$

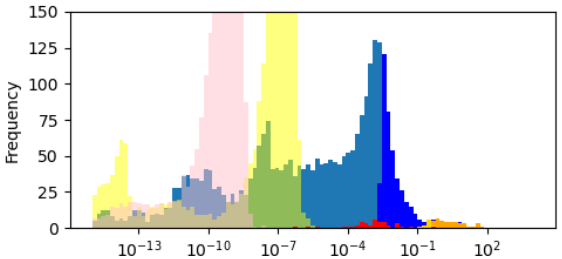
Eliminated space of $\tilde{H}^*_{C^*}$



Histogram of Absolute Eigenvalues of H^* and $\tilde{H}^*_{C^*}$

Effective space of H^*

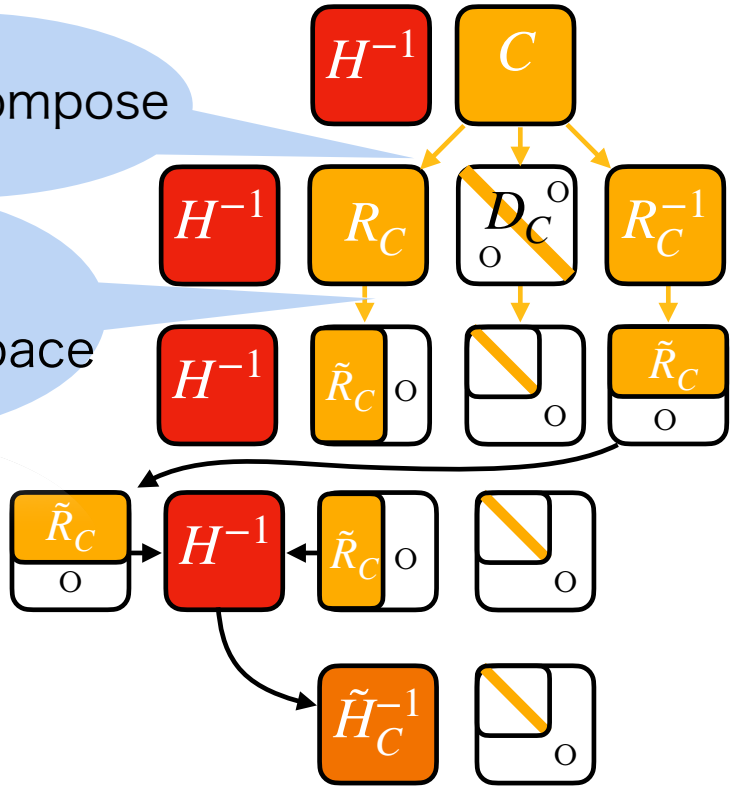
Eliminated space of H^*



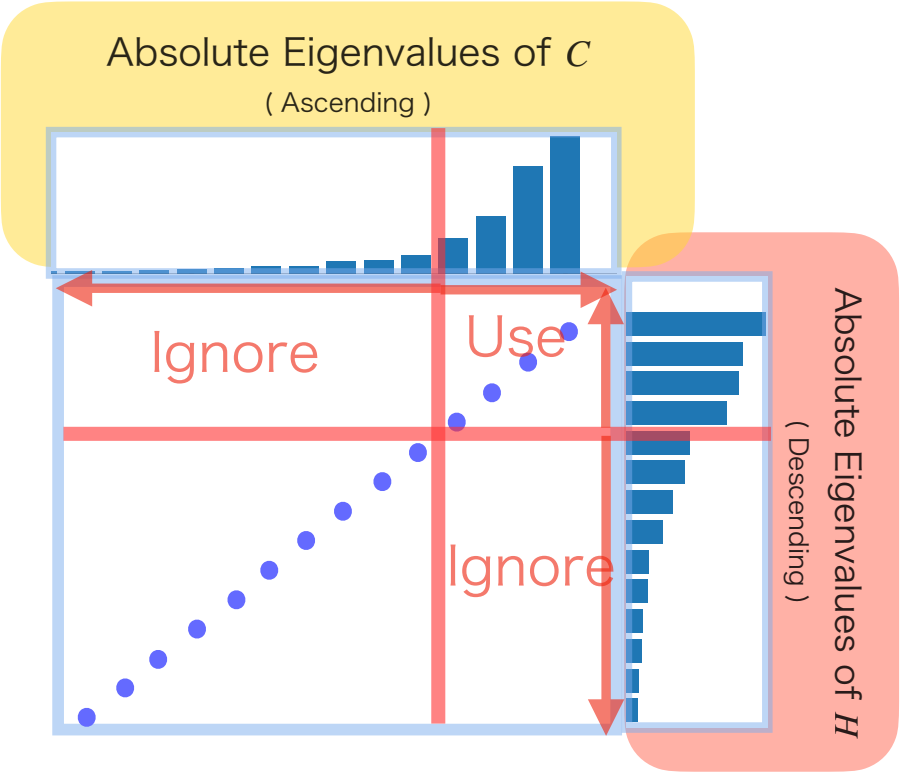
Histogram of Absolute Eigenvalues of C^*, H^* and $\tilde{H}^*_{C^*}$

Eigen decompose

Eliminate Degenerate space



What would happen if Eigenvalues of C^* were used?



Assumption

(because H^* and C^* are deviations from F)

- Order of magnitude of Eigenvalues should be preserved
  - So, accuracy of NIC^* shouldn't be affected

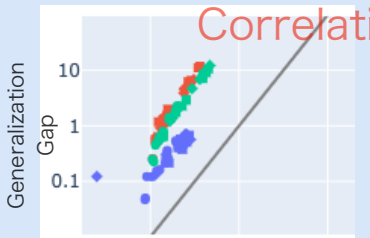
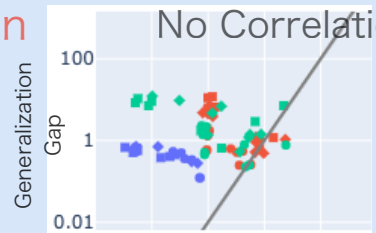
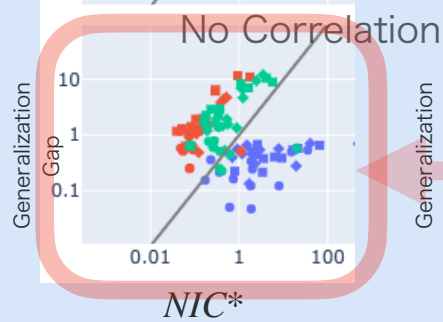
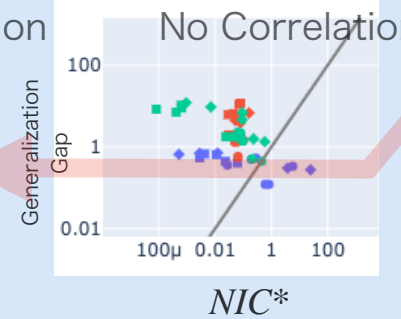
Then why use C^*?

- Distribution is separated ⋯ might be a more natural method to choose effective/ignored Eigenvalues

Estimating the Generalization Gap with NIC^*

Results

(Correlation→Successful, No correlation→unsuccessful)

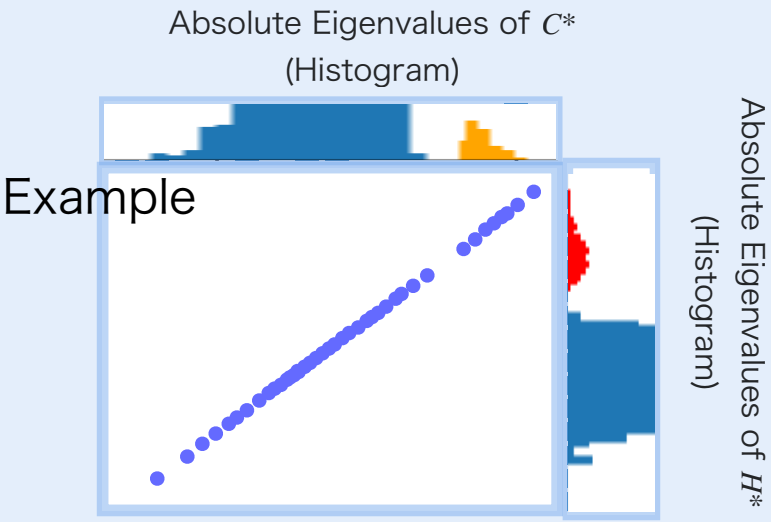
Select Eigenspaces based on…	Test dataset	Training dataset
H^* 【Past】		
C^* 【New】		

- Whether Eigenbases of C^* or H^* are used shouldn't affect NIC^* accuracy
  - There should be a correlation
- Degenerate spaces are not preserved?

What the results imply

If NIC^* performs identically between eigenspaces of H^* and C^*

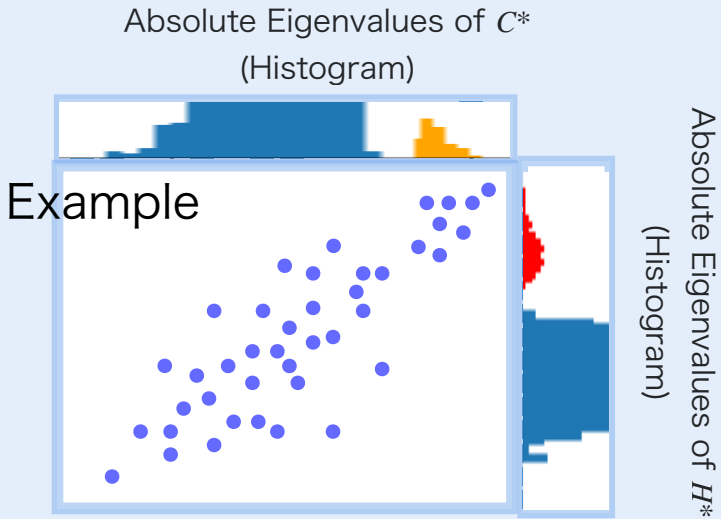
Order of magnitude of the eigenspaces between H^* and C^* are preserved



Scatter plot of Absolute Eigenvalues of C and H with corresponding eigenspaces

If NIC^* DOES NOT perform identically between eigenspaces of H^* and C^*

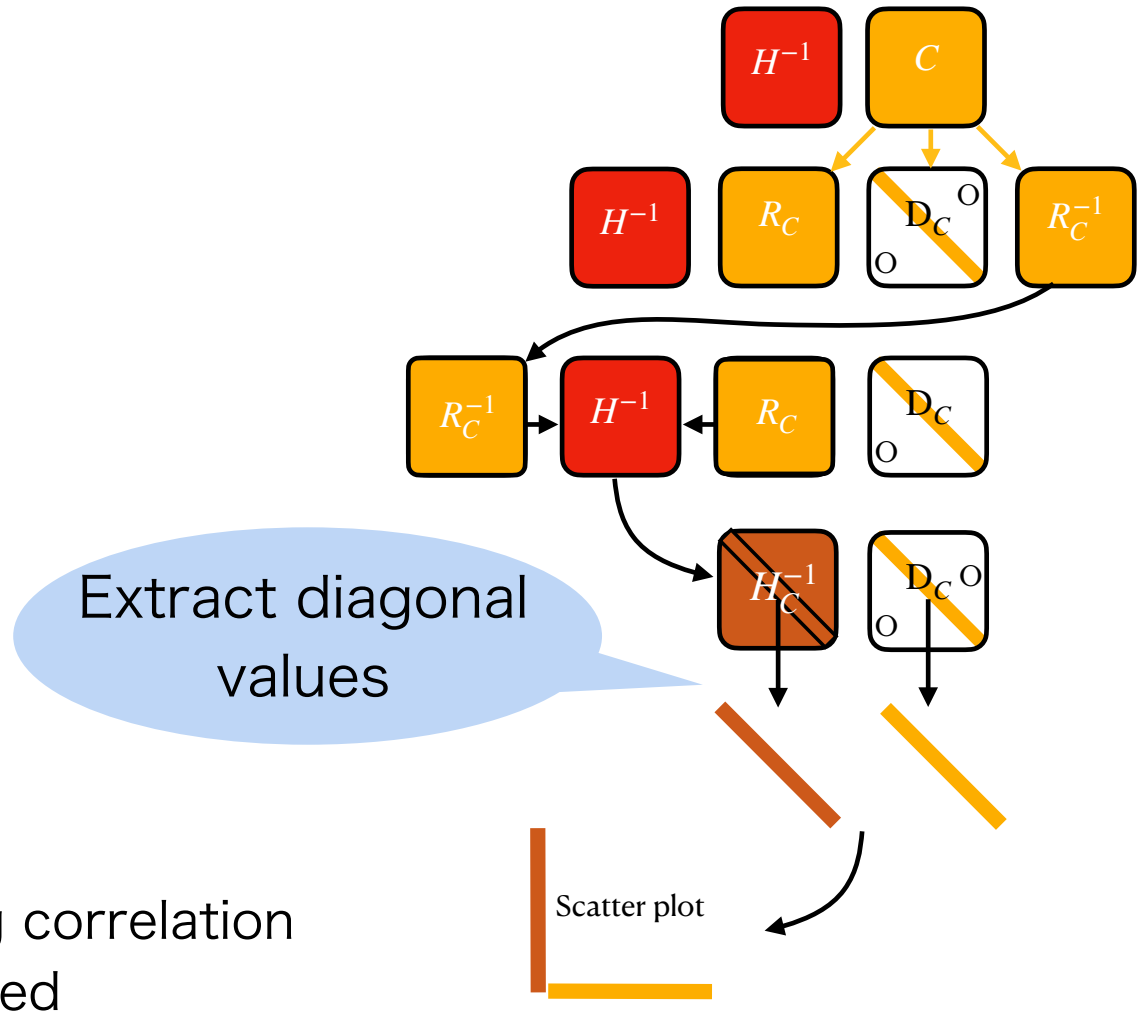
Order of magnitude of the eigenspaces between H^* and C^* are NOT preserved



…is this causing NIC^* to fail?

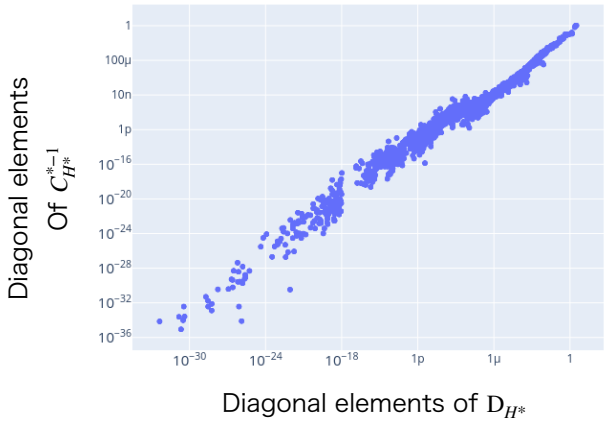
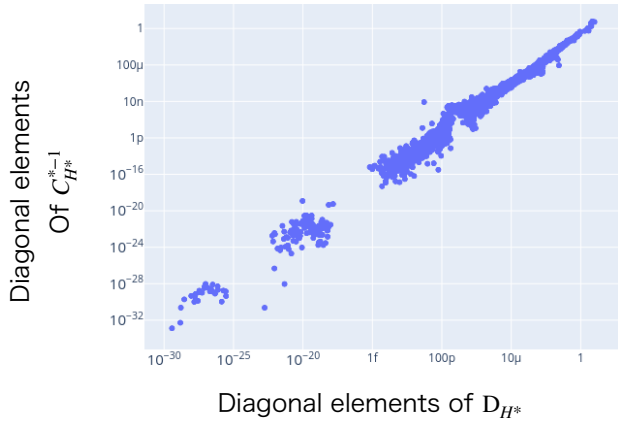
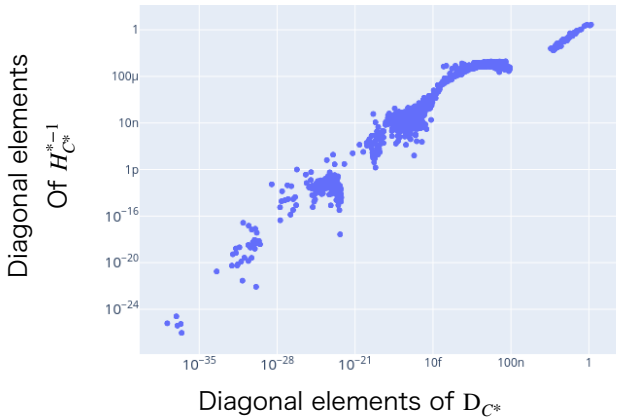
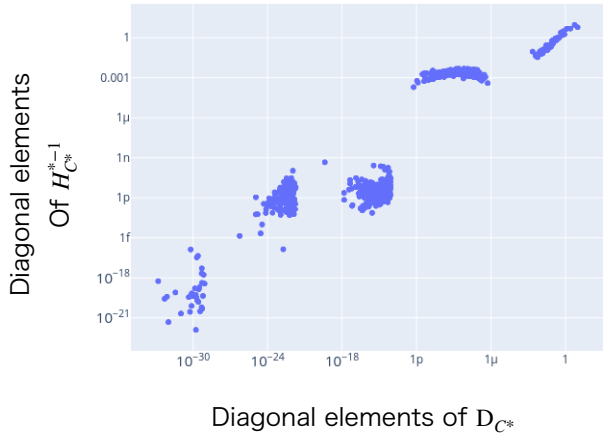
Transforming H with Eigenbases of C

$$
\begin{aligned}
\mathrm{tr}\left(H^{-1}C\right) &= \mathrm{tr}\left(H^{-1}\left(R_C D_C R_C^{-1}\right)\right) \\
&= \mathrm{tr}\left(H^{-1}R_C D_C R_C^{-1}\right) \\
&= \mathrm{tr}\left(R_C^{-1}H^{-1}R_C D_C\right) \\
&= \mathrm{tr}\left(\left(R_C H R_C^{-1}\right)^{-1} D_C\right) \\
&= \mathrm{tr}\left(H_C^{-1} D_C\right)
\end{aligned}
$$



- Scatter plot should show a strong correlation if order of magnitude are preserved

Comparing Diagonal spaces with shared Eigenbases

Shared Eigenbases	Model 1	Model 2
H^* (Transform C^* with Eigenbases of H^*)		
C^* (Transform H^* with Eigenbases of C^*)		

NIC is a generalization of the AIC

Known, used for empirical NIC

$H^*(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^*)$ $C^*(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^*)$

If $N \to \infty$, then $H = H^*, C = C^*$

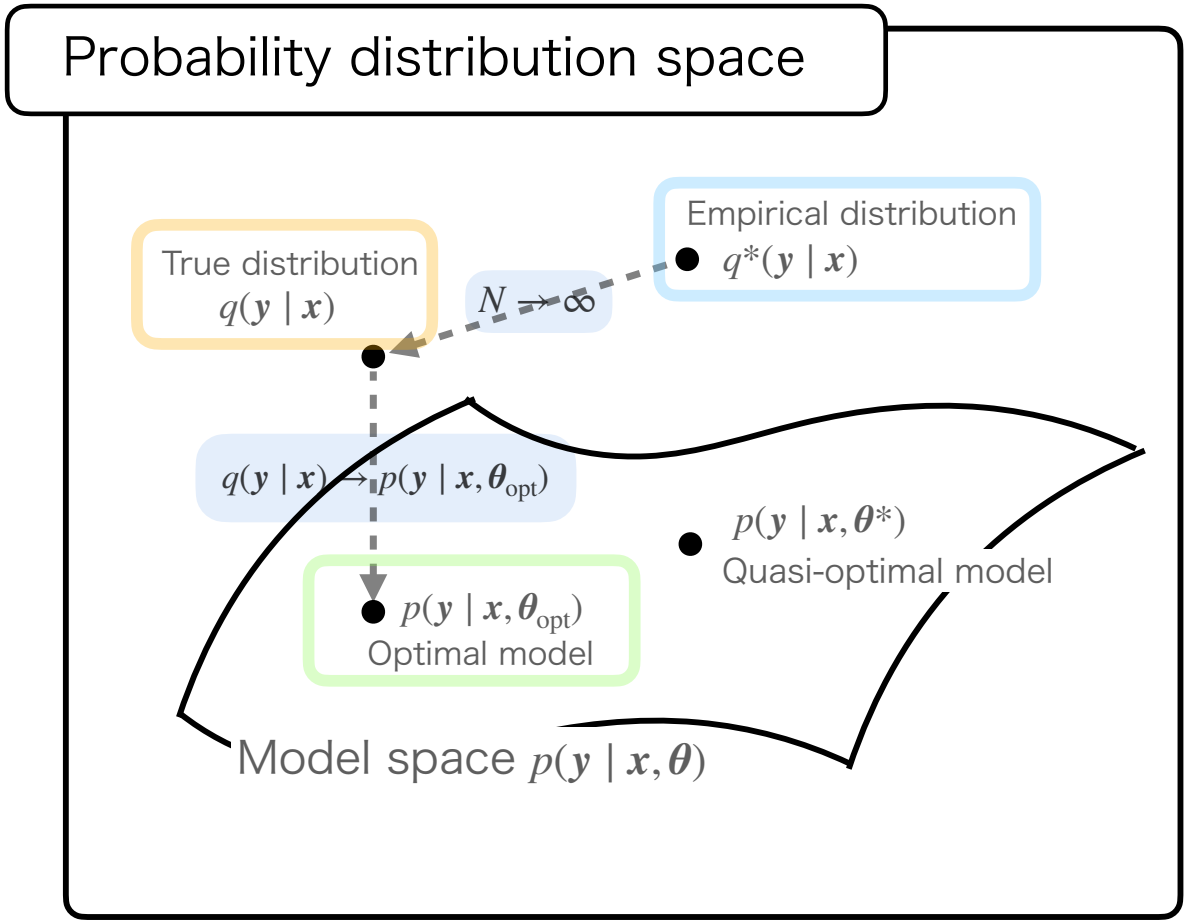
Unknown

$H(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}_{\mathrm{opt}})$ $C(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}_{\mathrm{opt}})$

If $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{opt}}) = q(\boldsymbol{y} \mid \boldsymbol{x})$, then $F = H = C$

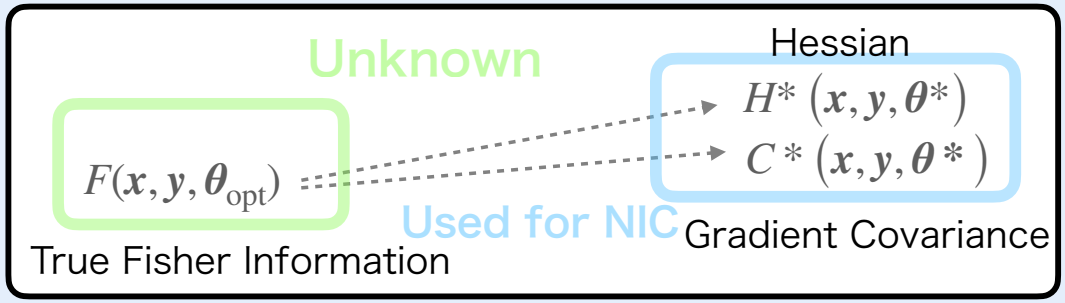
Unknown

$F(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}_{\mathrm{opt}})$



What does low correlation mean?

H and C are approximations of "true F"



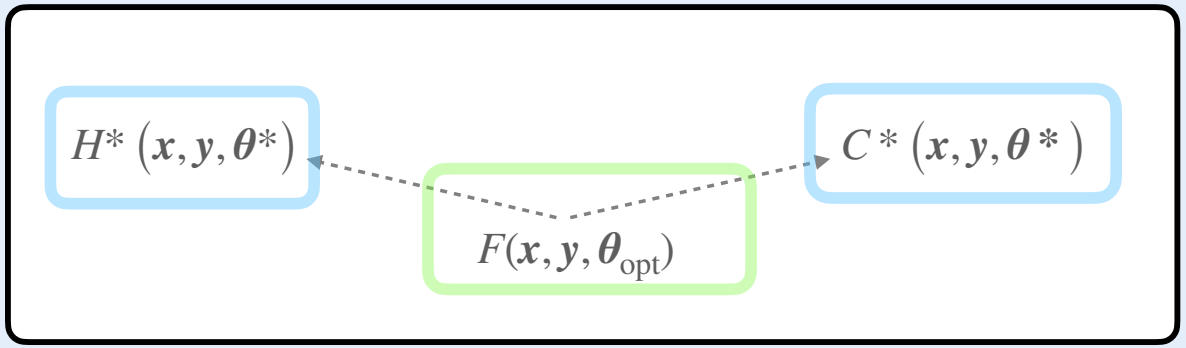
- When 2 conditions are met,

$$F(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}_{\mathrm{opt}}) = H^*(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^*) = C^*(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^*)$$

If low correlation

- Degenerate spaces are not preserved between H^* and C^*
  - Something is causing unexpectedly large fluctuations

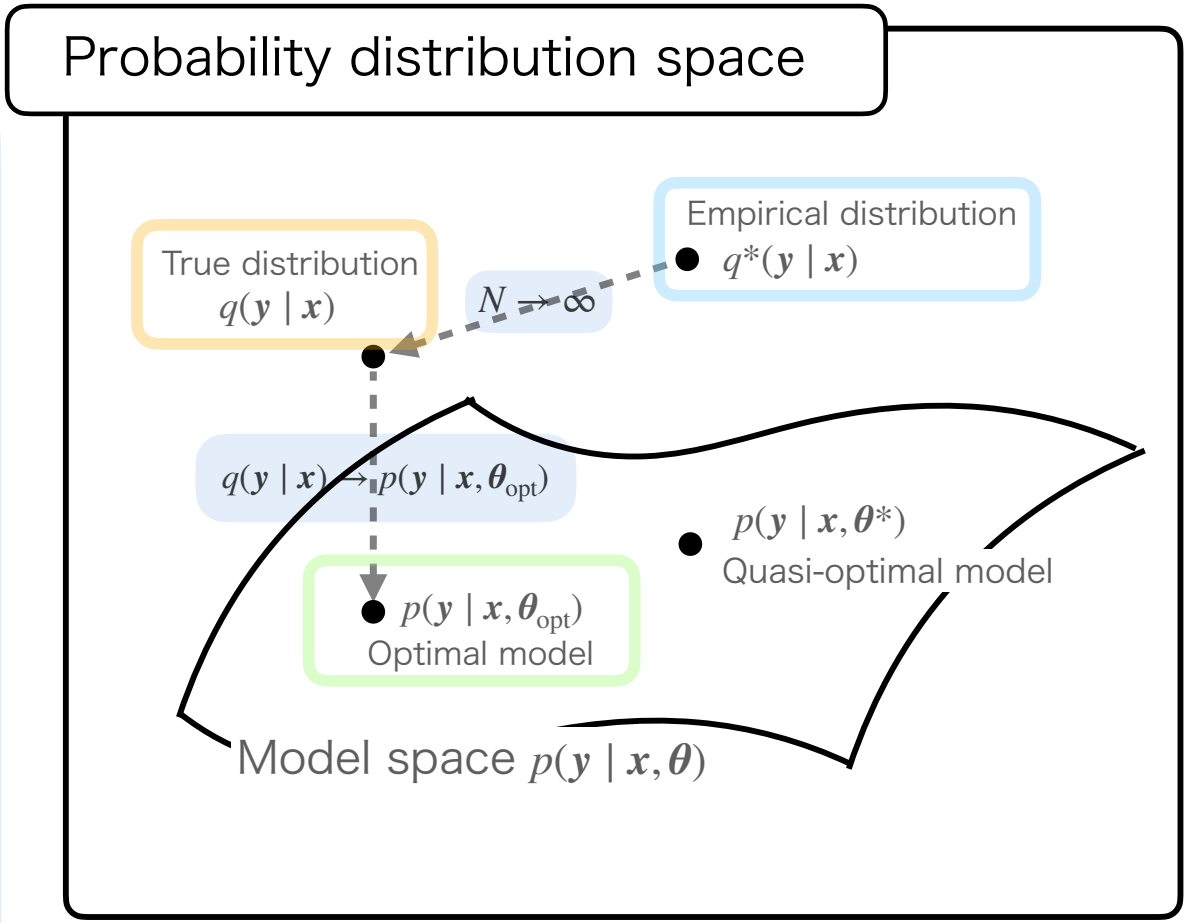
➡ Where?



New Question

What deviation causes bad estimation with NIC^*?

- In order to observe where the fluctuations occur, H^* and C^* has to be compared to F
- F is unobtainable for real-world dataset
- Should experiment with ideal conditions (artificial dataset)

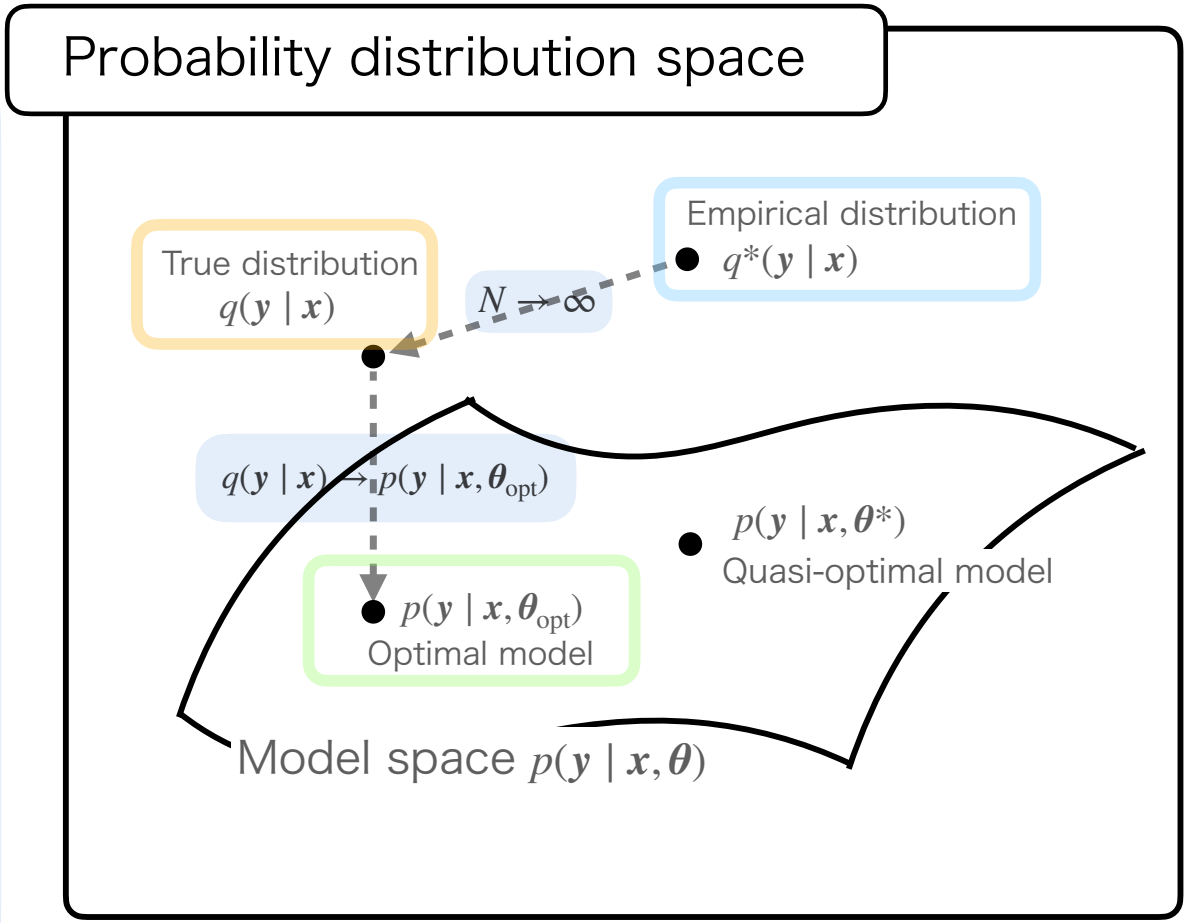


Conclusion and Future Work

- There is something causing fluctuations between H^* and C^*, for DNNs

Question: What deviation is the cause of this fluctuation?

- Reproduce experiment under ideal conditions, gradually deviate, and find where fluctuations occur



Appendix

Estimating Generalization Loss in a different way

- Generalization Loss > Train Loss (which was minimized)

  $D(q, p(\boldsymbol{\theta}^*))$ $D(q^*, p(\boldsymbol{\theta}^*))$

- Define: "Generalization Gap" ≡ Generalization Loss – Train Loss (> 0)

  $\mathscr{G}$ $D(q, p(\boldsymbol{\theta}^*))$ $D(q^*, p(\boldsymbol{\theta}^*))$

then,

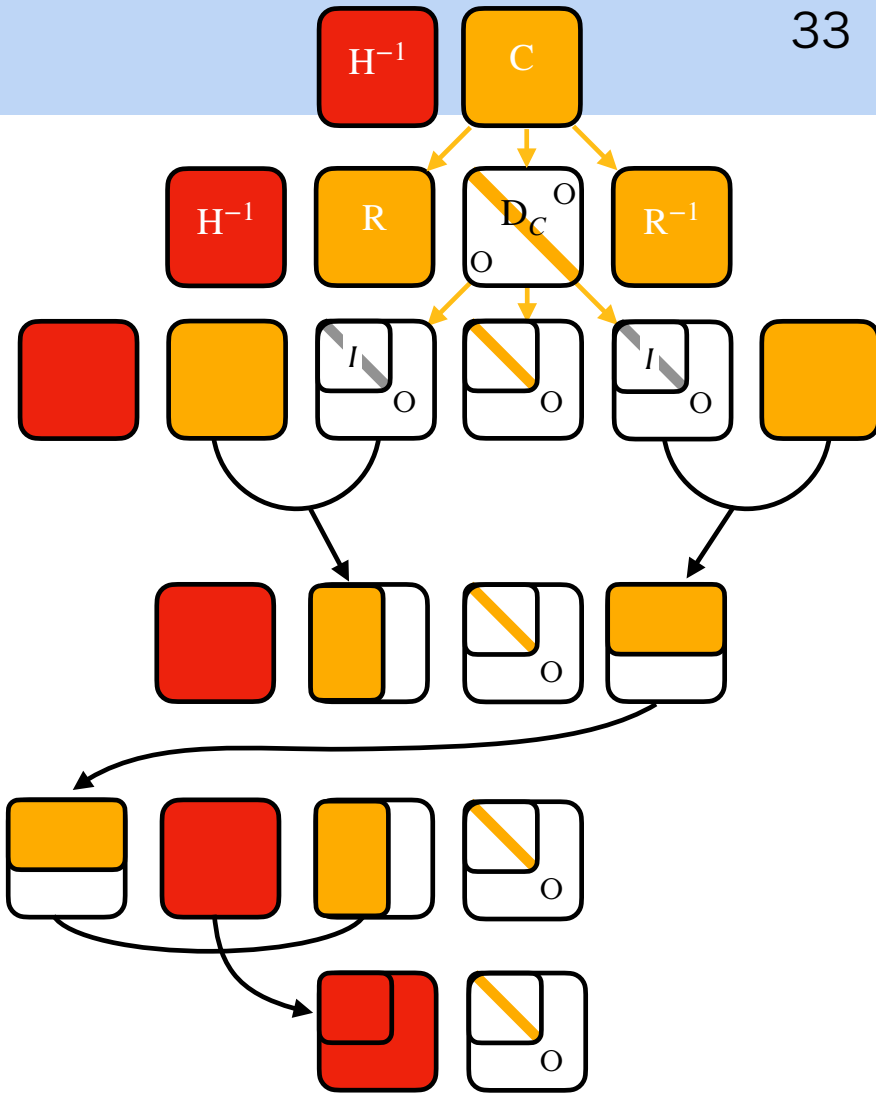
Generalization Loss = Generalization Gap + Train Loss

$$D(q, p(\boldsymbol{\theta}^*)) = \mathscr{G} + D(q^*, p(\boldsymbol{\theta}^*))$$

(unknown) (unknown) (known)

- If we know $\mathscr{G}$, we know the Generalization Loss

➡ Estimate the Generalization Gap? ➡ Information Criterion

Transforming H with Eigenspace of C

$$
\begin{aligned}
\operatorname{tr}\left(H^{-1}C\right) &= \operatorname{tr}\left(H^{-1}\left(R_C D_C R_C^{-1}\right)\right) \\
&= \operatorname{tr}\left(H^{-1}\left(R\tilde{I}_C D_C \tilde{I}_C^{-1} R^{-1}\right)\right) \\
&= \operatorname{tr}\left(H^{-1}\left(R\tilde{I}_C\right) D_C \left(\tilde{I}_C^{-1} R^{-1}\right)\right) \\
&= \operatorname{tr}\left(H^{-1}\tilde{R}_C D_C \tilde{R}_C^{-1}\right) \\
&= \operatorname{tr}\left(\tilde{R}_C^{-1} H^{-1} \tilde{R}_C D_C\right) \\
&= \operatorname{tr}\left(\left(\tilde{R}_C H \tilde{R}_C^{-1}\right)^{-1} D_C\right) \\
&= \operatorname{tr}\left(\tilde{H}_C^{-1} D_C\right)
\end{aligned}
$$