The Size of Teachers as a Measure of Data Complexity

PAC-Bayes Excess Risk Bounds and Scaling Laws for Neural Networks

Gintare Karolina Dziugaite (Google Deepmind)

Daniel M. Roy (U of Toronto; Vector Institute)

Post-Bayes Online Seminar

Generalization in deep learning depends on the data.

Generalization in deep learning depends on the data.

**We have no good way (yet) of modelling real data,
e.g., no notion of the complexity of data.**

Generalization in deep learning depends on the data.

**We have no good way (yet) of modelling real data,
e.g., no notion of the complexity of data.**

**OTOH, empirical “neural scaling laws” predict performance
across wide ranges of data and model sizes.**

Generalization in deep learning depends on the data.

**We have no good way (yet) of modelling real data,
e.g., no notion of the complexity of data.**

**OTOH, empirical “neural scaling laws” predict performance
across wide ranges of data and model sizes.**

What might neural scaling laws tell us about the complexity of the underlying data?

Evidence for Complexity: A Toy Theory

Evidence for Complexity: A Toy Theory

Let $r(\theta)$ be the risk (say, probability of misclassification) of a predictor θ.

Evidence for Complexity: A Toy Theory

Let $r(\theta)$ be the risk (say, probability of misclassification) of a predictor θ.

Let $\hat{\theta}_n$ be the predictor produced by our learning algorithm
- given n training data
- training a model of size $m(n)$.

Evidence for Complexity: A Toy Theory

Let $r(\theta)$ be the risk (say, probability of misclassification) of a predictor θ.

Let $\hat{\theta}_n$ be the predictor produced by our learning algorithm
- given n training data
- training a model of size $m(n)$.

Theorem. For every data distribution μ, there exists a function $C_\mu(\varepsilon)$ such that, with high probability under $\mu^{\otimes n}$,

$$r(\hat{\theta}_n) \leq \inf_{\varepsilon} \left\{ \varepsilon + O\left(\sqrt{\frac{\varepsilon \cdot C_\mu(\varepsilon)}{n}} \right) \right\}.$$

Evidence for Complexity: A Toy Theory

Let $r(\theta)$ be the risk (say, probability of misclassification) of a predictor θ.

Let $\hat{\theta}_n$ be the predictor produced by our learning algorithm
given n training data
training a model of size $m(n)$.

Theorem. For every data distribution μ, there exists a function $C_\mu(\varepsilon)$ such that, with high probability under $\mu^{\otimes n}$,

$$r(\hat{\theta}_n) \leq \inf_{\varepsilon} \left\{ \varepsilon + O\left(\sqrt{\frac{\varepsilon \cdot C_\mu(\varepsilon)}{n}} \right) \right\}.$$

The "complexity" function $C(\varepsilon)$ dictates risk rate:

$$\begin{aligned}
&\text{if } C_\mu(\varepsilon) = O(\text{polylog}(\varepsilon^{-1})) && \text{then } r(\hat{\theta}) = O(n^{-1}) \\
&\text{if } C_\mu(\varepsilon) = O(\varepsilon^{-p}) && \text{then } r(\hat{\theta}) = O(n^{-1/(p+1)}) \\
&\text{if } C_\mu(\varepsilon) = O(\exp(\text{poly}(\varepsilon^{-1}))) && \text{then } r(\hat{\theta}) = O(\log^{-1} n).
\end{aligned}$$

Evidence for Complexity: A Toy Theory

Let $r(\theta)$ be the risk (say, probability of misclassification) of a predictor θ.

Let $\hat{\theta}_n$ be the predictor produced by our learning algorithm
given n training data
training a model of size $m(n)$.

Theorem. For every data distribution μ, there exists a function $C_\mu(\varepsilon)$ such that, with high probability under $\mu^{\otimes n}$,

$$r(\hat{\theta}_n) \leq \inf_{\varepsilon} \left\{ \varepsilon + O\left(\sqrt{\frac{\varepsilon \cdot C_\mu(\varepsilon)}{n}} \right) \right\}.$$

The "complexity" function $C(\varepsilon)$ dictates risk rate:

$$\begin{aligned}
&\text{if } C_\mu(\varepsilon) = O(\text{polylog}(\varepsilon^{-1})) && \text{then } r(\hat{\theta}) = O(n^{-1}) \\
&\text{if } C_\mu(\varepsilon) = O(\varepsilon^{-p}) && \text{then } r(\hat{\theta}) = O(n^{-1/(p+1)}) \\
&\text{if } C_\mu(\varepsilon) = O(\exp(\text{poly}(\varepsilon^{-1}))) && \text{then } r(\hat{\theta}) = O(\log^{-1} n).
\end{aligned}$$

Consequence: If scaling law for n data and size $m(n)$ model says risk decays *slower* than $O(n^{-1/(p+1)})$, then...

Evidence for Complexity: A Toy Theory

Let $r(\theta)$ be the risk (say, probability of misclassification) of a predictor θ.

Let $\hat{\theta}_n$ be the predictor produced by our learning algorithm
given n training data
training a model of size $m(n)$.

Theorem. For every data distribution μ, there exists a function $C_\mu(\varepsilon)$ such that, with high probability under $\mu^{\otimes n}$,

$$r(\hat{\theta}_n) \leq \inf_{\varepsilon} \left\{ \varepsilon + O\left(\sqrt{\frac{\varepsilon \cdot C_\mu(\varepsilon)}{n}} \right) \right\}.$$

The "complexity" function $C(\varepsilon)$ dictates risk rate:

$$\begin{aligned}
&\text{if } C_\mu(\varepsilon) = O(\text{polylog}(\varepsilon^{-1})) && \text{then } r(\hat{\theta}) = O(n^{-1}) \\
&\text{if } C_\mu(\varepsilon) = O(\varepsilon^{-p}) && \text{then } r(\hat{\theta}) = O(n^{-1/(p+1)}) \\
&\text{if } C_\mu(\varepsilon) = O(\exp(\text{poly}(\varepsilon^{-1}))) && \text{then } r(\hat{\theta}) = O(\log^{-1} n).
\end{aligned}$$

Consequence: If scaling law for n data and size $m(n)$ model
says risk decays *slower* than $O(n^{-1/(p+1)})$, then... $C_\mu(\varepsilon) \notin O(\varepsilon^{-p})$.

Evidence for Complexity: A Toy Theory

Let $r(\theta)$ be the risk (say, probability of misclassification) of a predictor θ.

Let $\hat{\theta}_n$ be the predictor produced by our learning algorithm
given n training data
training a model of size $m(n)$.

Theorem. For every data distribution μ, there exists a function $C_\mu(\varepsilon)$ such that, with high probability under $\mu^{\otimes n}$,

$$r(\hat{\theta}_n) \leq \inf_{\varepsilon} \left\{ \varepsilon + O\left(\sqrt{\frac{\varepsilon \cdot C_\mu(\varepsilon)}{n}} \right) \right\}.$$

The "complexity" function $C(\varepsilon)$ dictates risk rate:

$$\begin{aligned}
&\text{if } C_\mu(\varepsilon) = O(\text{polylog}(\varepsilon^{-1})) && \text{then } r(\hat{\theta}) = O(n^{-1}) \\
&\text{if } C_\mu(\varepsilon) = O(\varepsilon^{-p}) && \text{then } r(\hat{\theta}) = O(n^{-1/(p+1)}) \\
&\text{if } C_\mu(\varepsilon) = O(\exp(\text{poly}(\varepsilon^{-1}))) && \text{then } r(\hat{\theta}) = O(\log^{-1} n).
\end{aligned}$$

Consequence: If scaling law for n data and size $m(n)$ model
says risk decays *slower* than $O(n^{-1/(p+1)})$, then... $C_\mu(\varepsilon) \notin O(\varepsilon^{-p})$.

1. Such function $C_\mu(\cdot)$ can be viewed as defining a notion of complexity of the data distribution μ.

Evidence for Complexity: A Toy Theory

Let $r(\theta)$ be the risk (say, probability of misclassification) of a predictor θ.

Let $\hat{\theta}_n$ be the predictor produced by our learning algorithm
given n training data
training a model of size $m(n)$.

Theorem. For every data distribution μ, there exists a function $C_\mu(\varepsilon)$ such that, with high probability under $\mu^{\otimes n}$,

$$r(\hat{\theta}_n) \leq \inf_{\varepsilon} \left\{ \varepsilon + O\left(\sqrt{\frac{\varepsilon \cdot C_\mu(\varepsilon)}{n}} \right) \right\}.$$

The "complexity" function $C(\varepsilon)$ dictates risk rate:

$$\begin{aligned}
&\text{if } C_\mu(\varepsilon) = O(\text{polylog}(\varepsilon^{-1})) && \text{then } r(\hat{\theta}) = O(n^{-1}) \\
&\text{if } C_\mu(\varepsilon) = O(\varepsilon^{-p}) && \text{then } r(\hat{\theta}) = O(n^{-1/(p+1)}) \\
&\text{if } C_\mu(\varepsilon) = O(\exp(\text{poly}(\varepsilon^{-1}))) && \text{then } r(\hat{\theta}) = O(\log^{-1} n).
\end{aligned}$$

Consequence: If scaling law for n data and size $m(n)$ model
says risk decays *slower* than $O(n^{-1/(p+1)})$, then... $C_\mu(\varepsilon) \notin O(\varepsilon^{-p})$.

1. Such function $C_\mu(\cdot)$ can be viewed as defining a notion of complexity of the data distribution μ.
2. Scaling laws can provide evidence of *complexity*, in view of upper bounds.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Our Approach

We propose to measure the complexity of data in terms of ...

- the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.
- Can be viewed as a rate–distortion function, specific to the data distribution
  (Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

Our Approach

We propose to measure the complexity of data in terms of ...

- the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.
- Can be viewed as a rate–distortion function, specific to the data distribution
  (Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some “teacher” network has zero risk. (“realizable” setting)

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some “teacher” network has zero risk. (“realizable” setting)
  - **Conclusion (Buzaglo et al. ’24):** Number of teacher parameters determines sample complexity.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some “teacher” network has zero risk. (“realizable” setting)
  - **Conclusion (Buzaglo et al. ’24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We’ll quantize the weights (or need to introduce some notion of margin.)

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some “teacher” network has zero risk. (“realizable” setting)
  - **Conclusion (Buzaglo et al. ’24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We’ll quantize the weights (or need to introduce some notion of margin.)
- Primi

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathsf{d}\pi(\theta)\}$.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathsf{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.
- Secondi

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.
- Secondi
  - Turn agnostic bound into an oracle inequality:
    Risk of Gibbs posterior sample no more than any teacher's risk plus a size penalty.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.
- Secondi
  - Turn agnostic bound into an oracle inequality:
    Risk of Gibbs posterior sample no more than any teacher's risk plus a size penalty.
  - Introduce $C(\varepsilon)$ and rewrite bound, obtain $\inf_\varepsilon\{\varepsilon + \ldots . C(\varepsilon) \ldots\}$ bound.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.
- Secondi
  - Turn agnostic bound into an oracle inequality:
    Risk of Gibbs posterior sample no more than any teacher's risk plus a size penalty.
  - Introduce $C(\varepsilon)$ and rewrite bound, obtain $\inf_\varepsilon \{\varepsilon + \ldots . C(\varepsilon) \ldots\}$ bound.
  - **Conclusion:** Scaling laws suggest $C(\varepsilon) \in \omega(\varepsilon^{-p})$ for some p.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.
- Secondi
  - Turn agnostic bound into an oracle inequality:
    Risk of Gibbs posterior sample no more than any teacher's risk plus a size penalty.
  - Introduce $C(\varepsilon)$ and rewrite bound, obtain $\inf_\varepsilon \{\varepsilon + \ldots . C(\varepsilon) \ldots\}$ bound.
  - **Conclusion:** Scaling laws suggest $C(\varepsilon) \in \omega(\varepsilon^{-p})$ for some p.
- Dolce

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.
- Secondi
  - Turn agnostic bound into an oracle inequality:
    Risk of Gibbs posterior sample no more than any teacher's risk plus a size penalty.
  - Introduce $C(\varepsilon)$ and rewrite bound, obtain $\inf_\varepsilon \{\varepsilon + \ldots . C(\varepsilon) \ldots\}$ bound.
  - **Conclusion:** Scaling laws suggest $C(\varepsilon) \in \omega(\varepsilon^{-p})$ for some p.
- Dolce
  - Don't want teachers but *distributions on teachers*.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.
- Secondi
  - Turn agnostic bound into an oracle inequality:
    Risk of Gibbs posterior sample no more than any teacher's risk plus a size penalty.
  - Introduce $C(\varepsilon)$ and rewrite bound, obtain $\inf_\varepsilon \{\varepsilon + \ldots . C(\varepsilon) \ldots\}$ bound.
  - **Conclusion:** Scaling laws suggest $C(\varepsilon) \in \omega(\varepsilon^{-p})$ for some p.
- Dolce
  - Don't want teachers but *distributions on teachers*.
  - Distributions avoid quantization/discretization. (Bits back encoding.) BONUS: New perspective on variational inference.

Our Approach

We propose to measure the complexity of data in terms of ...

the size $C_\mu(\varepsilon)$ of the smallest network that achieves ε risk under μ.

Can be viewed as a rate–distortion function, specific to the data distribution

(Hafez-Kolahi, Moniri, Kasaei 2024; Hafez-Kolahi, Moniri et al. 2021)

- Antipasti
  - Study toy model of neural network training: *a randomly initialized neural networks that fit the data.*
  - Assume that some "teacher" network has zero risk. ("realizable" setting)
  - **Conclusion (Buzaglo et al. '24):** Number of teacher parameters determines sample complexity.
  - **Caveat:** We'll quantize the weights (or need to introduce some notion of margin.)
- Primi
  - Study less toy model: *sample from the Gibbs posterior* $\propto \exp\{-\beta \hat{r}_n(\theta) + \mathrm{d}\pi(\theta)\}$.
  - Drop assumption that some teacher network has zero risk. ("agnostic" setting)
  - **Conclusion:** Number of teacher parameters determines sample complexity (for *excess* risk).
  - **Bonus:** Nonvacuous bounds for MNIST.
- Secondi
  - Turn agnostic bound into an oracle inequality:
    Risk of Gibbs posterior sample no more than any teacher's risk plus a size penalty.
  - Introduce $C(\varepsilon)$ and rewrite bound, obtain $\inf_\varepsilon \{\varepsilon + \ldots . C(\varepsilon) \ldots\}$ bound.
  - **Conclusion:** Scaling laws suggest $C(\varepsilon) \in \omega(\varepsilon^{-p})$ for some p.
- Dolce
  - Don't want teachers but *distributions on teachers*.
  - Distributions avoid quantization/discretization. (Bits back encoding.) BONUS: New perspective on variational inference.
  - **Conclusion:** $C(\varepsilon)$ more complicated.

Learning Setup

Labelled data: denoted by $z, z_i, Z, \ldots$

Predictors: identified with parameters, denoted by $\theta, \hat{\theta}, \ldots$

Loss: $\ell(\theta, z)$, e.g., $\ell(\theta, (x, y)) = \mathbb{I}(\text{network with weights } \theta \text{ on input } x \text{ does not output label } y)$

Data distribution: μ, presumed unknown

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Prior: distribution π on Θ, which is nonrandom.

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Let $\hat{\Theta}_0 = \{\theta \in \Theta : \hat{r}_n(\theta) = 0\}$ be the set of interpolating predictors (i.e., with zero empirical risk).

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Let $\hat{\Theta}_0 = \{\theta \in \Theta : \hat{r}_n(\theta) = 0\}$ be the set of interpolating predictors (i.e., with zero empirical risk). Note: $\hat{\Theta}_0 \supseteq \Theta^*$.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Let $\hat{\Theta}_0 = \{\theta \in \Theta : \hat{r}_n(\theta) = 0\}$ be the set of interpolating predictors (i.e., with zero empirical risk). Note: $\hat{\Theta}_0 \supseteq \Theta^*$.

We want a random element from $\hat{\Theta}_0$. How?

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Let $\hat{\Theta}_0 = \{\theta \in \Theta : \hat{r}_n(\theta) = 0\}$ be the set of interpolating predictors (i.e., with zero empirical risk). Note: $\hat{\Theta}_0 \supseteq \Theta^*$.

We want a random element from $\hat{\Theta}_0$. How?

Fix a probability measure π on Θ.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Let $\hat{\Theta}_0 = \{\theta \in \Theta : \hat{r}_n(\theta) = 0\}$ be the set of interpolating predictors (i.e., with zero empirical risk). Note: $\hat{\Theta}_0 \supseteq \Theta^*$.

We want a random element from $\hat{\Theta}_0$. How?

Fix a probability measure π on Θ. Let $\theta \sim \pi$. If $\theta \in \hat{\Theta}_0$, accept and set $\hat{\theta} = \theta$. Otherwise, try again.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Let $\hat{\Theta}_0 = \{\theta \in \Theta : \hat{r}_n(\theta) = 0\}$ be the set of interpolating predictors (i.e., with zero empirical risk). Note: $\hat{\Theta}_0 \supseteq \Theta^*$.

We want a random element from $\hat{\Theta}_0$. How?

Fix a probability measure π on Θ. Let $\theta \sim \pi$. If $\theta \in \hat{\Theta}_0$, accept and set $\hat{\theta} = \theta$. Otherwise, try again.

Distribution of $\hat{\theta}$ is the "posterior" $\hat{\rho} = \pi(\,\cdot \mid \hat{\Theta}_0)$ given by $\hat{\rho}(A) = \dfrac{\pi(A \cap \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Let $\hat{\Theta}_0 = \{\theta \in \Theta : \hat{r}_n(\theta) = 0\}$ be the set of interpolating predictors (i.e., with zero empirical risk). Note: $\hat{\Theta}_0 \supseteq \Theta^*$.

We want a random element from $\hat{\Theta}_0$. How?

Fix a probability measure π on Θ. Let $\theta \sim \pi$. If $\theta \in \hat{\Theta}_0$, accept and set $\hat{\theta} = \theta$. Otherwise, try again.

Distribution of $\hat{\theta}$ is the "posterior" $\hat{\rho} = \pi(\,\cdot \mid \hat{\Theta}_0)$ given by $\hat{\rho}(A) = \dfrac{\pi(A \cap \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$. Equivalently, $\dfrac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \dfrac{\mathbb{I}(\theta \in \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$.

Antipasti or prior work by Buzaglo, Harel, Nacson, Brutzkus, Srebro, Soudry (2024)

Buzaglo et al.'s assumptions

Let $\Theta^* = \{\theta \in \Theta : r(\theta^*) = 0\}$.

Assumption (realizability). There exists $\theta^* \in \Theta^*$.
Assumption (finiteness). Θ is a finite set.

Buzaglo et al.: What's the risk of a "random interpolator"?

Let $\hat{\Theta}_0 = \{\theta \in \Theta : \hat{r}_n(\theta) = 0\}$ be the set of interpolating predictors (i.e., with zero empirical risk). Note: $\hat{\Theta}_0 \supseteq \Theta^*$.

We want a random element from $\hat{\Theta}_0$. How?

Fix a probability measure π on Θ. Let $\theta \sim \pi$. If $\theta \in \hat{\Theta}_0$, accept and set $\hat{\theta} = \theta$. Otherwise, try again.

Distribution of $\hat{\theta}$ is the "posterior" $\hat{\rho} = \pi(\,\cdot \mid \hat{\Theta}_0)$ given by $\hat{\rho}(A) = \dfrac{\pi(A \cap \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$. Equivalently, $\dfrac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \dfrac{\mathbb{I}(\theta \in \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$.

So... what's the risk of $\hat{\theta}$ sampled from the posterior $\hat{\rho}$?

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Let $\hat{\theta}$ be a sample from $\hat{\rho}$. (That is, $\hat{\theta} \mid S \sim \hat{\rho}$.)

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Let $\hat{\theta}$ be a sample from $\hat{\rho}$. (That is, $\hat{\theta} \mid S \sim \hat{\rho}$.)

PAC-Bayes offers comparisons for the random variables

$$r(\hat{\theta}) \quad \text{and} \quad \hat{r}_n(\hat{\theta}) \quad \text{and} \quad \underbrace{\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta})}_{\text{information density}}$$

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Let $\hat{\theta}$ be a sample from $\hat{\rho}$. (That is, $\hat{\theta} \mid S \sim \hat{\rho}$.)

PAC-Bayes offers comparisons for the random variables

$$r(\hat{\theta}) \quad \text{and} \quad \hat{r}_n(\hat{\theta}) \quad \text{and} \quad \underbrace{\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta})}_{\text{information density}}$$

Classical case controls expectations of these quantities under $\hat{\rho}$.

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Let $\hat{\theta}$ be a sample from $\hat{\rho}$. (That is, $\hat{\theta} \mid S \sim \hat{\rho}$.)

PAC-Bayes offers comparisons for the random variables

$$r(\hat{\theta}) \quad \text{and} \quad \hat{r}_n(\hat{\theta}) \quad \text{and} \quad \underbrace{\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta})}_{\text{information density}}$$

Classical case controls expectations of these quantities under $\hat{\rho}$.

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Theorem (Single sample bound; Catoni 2007)

Fix $\lambda > 0$. With probability at least $1 - \delta$,

$$r(\hat{\theta}) \leq \Phi^{-1}_{\lambda/n}\left(\hat{r}_n(\hat{\theta}) + \frac{\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta}) + \log \frac{1}{\delta}}{\lambda}\right) \qquad \textit{where } \Phi_a^{-1}(q) = (1 - \exp(-aq))/(1 - \exp(-a)).$$

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Theorem (Single sample bound; Catoni 2007)

Fix $\lambda > 0$. With probability at least $1 - \delta$,

$$r(\hat{\theta}) \leq \Phi^{-1}_{\lambda/n}\left(\hat{r}_n(\hat{\theta}) + \frac{\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta}) + \log \frac{1}{\delta}}{\lambda}\right) \qquad \textit{where } \Phi_a^{-1}(q) = (1 - \exp(-aq))/(1 - \exp(-a)).$$

How should we interpret $\Phi^{-1}_{\lambda/n}$? $\inf_{\lambda \in \mathbb{R}_+} \Phi^{-1}_{\lambda/n}\left(q + \frac{d}{\lambda}\right) \leq q + \frac{2d}{n} + \sqrt{\frac{2dq}{n}}$, provided r.h.s. is less than $1/2$.

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

Prior: distribution π on Θ, which is nonrandom.

Corollary

Assume $\hat{r}_n(\hat{\theta}) = 0$ *a.s. under* $\hat{\theta} \mid S \sim \hat{\rho}$.

Background: PAC-Bayes Framework

Risk: $r(\theta) = \int \ell(\theta, z)\mu(\mathrm{d}z)$

Data: $S = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$

Empirical Risk: $\hat{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}, z_i)$

Posterior: distribution $\hat{\rho}$ on Θ, which may depend on S.

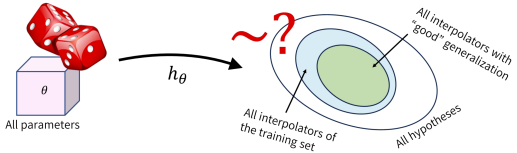
Prior: distribution π on Θ, which is nonrandom.

Corollary

Assume $\hat{r}_n(\hat{\theta}) = 0$ *a.s. under* $\hat{\theta} \mid S \sim \hat{\rho}$. *With probability at least* $1 - \delta$,

$$r(\hat{\theta}) \leq \frac{\log \frac{d\hat{\rho}}{d\pi}(\hat{\theta}) + \log \frac{1}{\delta}}{n}$$

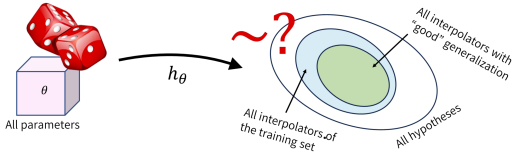
Antipasti with a side of PAC-Bayes



Antipasti with a side of PAC-Bayes

Recall: $\hat{\rho}$ is posterior given interpolation, given by

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{\mathbb{I}(\theta \in \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$$



Antipasti with a side of PAC-Bayes

Recall: $\hat{\rho}$ is posterior given interpolation, given by

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{\mathbb{I}(\theta \in \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$$

.

By construction, $\hat{r}_n(\theta) = 0$ and $\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \log \frac{1}{\pi(\Theta^*)}$ for $\hat{\rho}$-almost all θ.

Antipasti with a side of PAC-Bayes

Recall: $\hat{\rho}$ is posterior given interpolation, given by

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{\mathbb{I}(\theta \in \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$$

.

By construction, $\hat{r}_n(\theta) = 0$ and $\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \log \frac{1}{\pi(\Theta^*)}$ for $\hat{\rho}$-almost all θ.

What's the risk of $\hat{\theta}$ sampled from $\hat{\rho}$?

Antipasti with a side of PAC-Bayes

Recall: $\hat{\rho}$ is posterior given interpolation, given by

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{\mathbb{I}(\theta \in \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)} \quad .$$

By construction, $\hat{r}_n(\theta) = 0$ and $\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \log \frac{1}{\pi(\Theta^*)}$ for $\hat{\rho}$-almost all θ.

What's the risk of $\hat{\theta}$ sampled from $\hat{\rho}$?

Lemma (Risk of posterior sampling; Dziugaite & R. 2025)

Antipasti with a side of PAC-Bayes

Recall: $\hat{\rho}$ is posterior given interpolation, given by

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{\mathbb{I}(\theta \in \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)} \quad .$$

By construction, $\hat{r}_n(\theta) = 0$ and $\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \log \frac{1}{\pi(\Theta^*)}$ for $\hat{\rho}$-almost all θ.

What's the risk of $\hat{\theta}$ sampled from $\hat{\rho}$?

Lemma (Risk of posterior sampling; Dziugaite & R. 2025)

Fix π and assume π-realizability. Let $\hat{\theta} \mid S \sim \hat{\rho}$. With probability at least $1 - \delta$,

Antipasti with a side of PAC-Bayes

Recall: $\hat{\rho}$ is posterior given interpolation, given by

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{\mathbb{I}(\theta \in \hat{\Theta}_0)}{\pi(\hat{\Theta}_0)}$$

.

By construction, $\hat{r}_n(\theta) = 0$ and $\log \frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \log \frac{1}{\pi(\Theta^*)}$ for $\hat{\rho}$-almost all θ.

What's the risk of $\hat{\theta}$ sampled from $\hat{\rho}$?

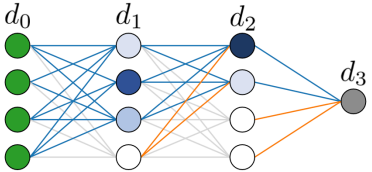
Lemma (Risk of posterior sampling; Dziugaite & R. 2025)

Fix π and assume π-realizability. Let $\hat{\theta} \mid S \sim \hat{\rho}$. With probability at least $1 - \delta$,

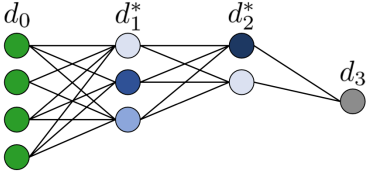
$$r(\hat{\theta}) \leq \frac{\log \frac{1}{\pi(\Theta^*)} + \log \frac{1}{\delta}}{n}$$

Antipasti: Buzaglo et al.'s lower bound on $\pi(\Theta^*)$, the probability of interpolating

For a network θ, let $m(\theta)$ be the smallest number of parameters we must specify to produce a network functionally equivalent to θ.



student architecture

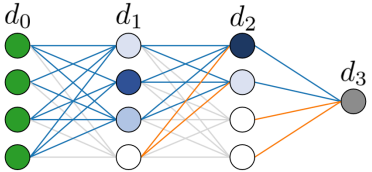


teacher architecture

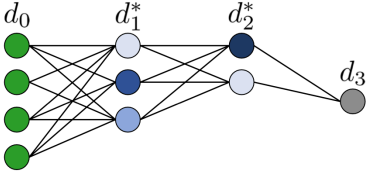
Note: $m(\theta) =$ # parameters in teacher network + # neurons in student network

Antipasti: Buzaglo et al.'s lower bound on $\pi(\Theta^*)$, the probability of interpolating

For a network θ, let $m(\theta)$ be the smallest number of parameters we must specify to produce a network functionally equivalent to θ.



student architecture



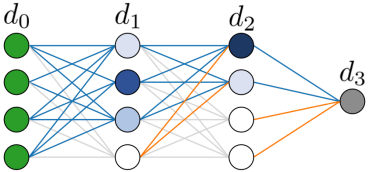
teacher architecture

Note: $m(\theta) =$ # parameters in teacher network + # neurons in student network

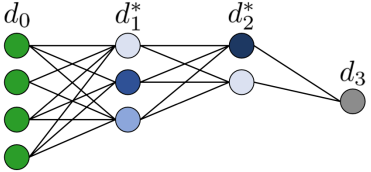
Assumption (quantization). Weights of all networks are quantized into Q levels, and zero is one level.

Antipasti: Buzaglo et al.'s lower bound on $\pi(\Theta^*)$, the probability of interpolating

For a network θ, let $m(\theta)$ be the smallest number of parameters we must specify to produce a network functionally equivalent to θ.



student architecture



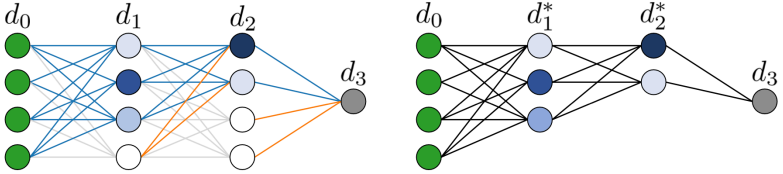
teacher architecture

Note: $m(\theta) =$ # parameters in teacher network + # neurons in student network

Assumption (quantization). Weights of all networks are quantized into Q levels, and zero is one level.
Assumption (batch-norm-like init). Weights leaving each neuron are multiplied by a neuron-specific scaling weight.

Antipasti: Buzaglo et al.'s lower bound on $\pi(\Theta^*)$, the probability of interpolating

For a network θ, let $m(\theta)$ be the smallest number of parameters we must specify to produce a network functionally equivalent to θ.



student architecture

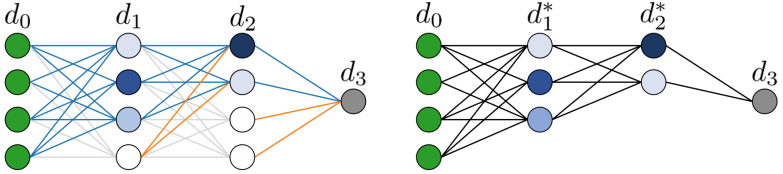
teacher architecture

Note: $m(\theta) =$ # parameters in teacher network + # neurons in student network

Assumption (quantization). Weights of all networks are quantized into Q levels, and zero is one level.
Assumption (batch-norm-like init). Weights leaving each neuron are multiplied by a neuron-specific scaling weight.
Assumption (uniform prior). π is uniform distribution over all quantized student networks.

Antipasti: Buzaglo et al.'s lower bound on $\pi(\Theta^*)$, the probability of interpolating

For a network θ, let $m(\theta)$ be the smallest number of parameters we must specify to produce a network functionally equivalent to θ.



student architecture

teacher architecture

Note: $m(\theta) =$ # parameters in teacher network + # neurons in student network

Assumption (quantization). Weights of all networks are quantized into Q levels, and zero is one level.
Assumption (batch-norm-like init). Weights leaving each neuron are multiplied by a neuron-specific scaling weight.
Assumption (uniform prior). π is uniform distribution over all quantized student networks.

Second contribution of Buzaglo et al.

Let $\theta^* \in \Theta^*$. Then $\log \frac{1}{\pi(\Theta^*)} = O(m(\theta^*) \log Q)$.

Last bite of Antipasti: Final conclusion of Buzaglo et al.

Last bite of Antipasti: Final conclusion of Buzaglo et al.

Theorem (Buzaglo et al. 2024; Dziugaite & R. 2025)

Last bite of Antipasti: Final conclusion of Buzaglo et al.

Theorem (Buzaglo et al. 2024; Dziugaite & R. 2025)

Assume we have quantization, batch-norm-like init, a uniform prior π, and π-realizability.

Last bite of Antipasti: Final conclusion of Buzaglo et al.

Theorem (Buzaglo et al. 2024; Dziugaite & R. 2025)

Assume we have quantization, batch-norm-like init, a uniform prior π, and π-realizability.

Let $m^* = \inf_{\theta^* \in \Theta^*} m(\theta^*)$ be the smallest teacher network in Θ^*.

Last bite of Antipasti: Final conclusion of Buzaglo et al.

Theorem (Buzaglo et al. 2024; Dziugaite & R. 2025)

Assume we have quantization, batch-norm-like init, a uniform prior π, and π-realizability.

Let $m^* = \inf_{\theta^* \in \Theta^*} m(\theta^*)$ be the smallest teacher network in Θ^*.

Let S be n i.i.d. data and let $\hat{\theta}$ be a random interpolating network.

Last bite of Antipasti: Final conclusion of Buzaglo et al.

Theorem (Buzaglo et al. 2024; Dziugaite & R. 2025)

Assume we have quantization, batch-norm-like init, a uniform prior π, and π-realizability.

Let $m^* = \inf_{\theta^* \in \Theta^*} m(\theta^*)$ be the smallest teacher network in Θ^*.

Let S be n i.i.d. data and let $\hat{\theta}$ be a random interpolating network.
With probability at least $1 - \delta$,

Last bite of Antipasti: Final conclusion of Buzaglo et al.

Theorem (Buzaglo et al. 2024; Dziugaite & R. 2025)

Assume we have quantization, batch-norm-like init, a uniform prior π, and π-realizability.

Let $m^* = \inf_{\theta^* \in \Theta^*} m(\theta^*)$ be the smallest teacher network in Θ^*.

Let S be n i.i.d. data and let $\hat{\theta}$ be a random interpolating network.
With probability at least $1 - \delta$,

$$r(\hat{\theta}) \leq \frac{m^* \log Q + \log \frac{1}{\delta}}{n}$$

Equivalently, if we have at least

$$n \geq \frac{m^* \log Q + \log \frac{1}{\delta}}{\varepsilon}$$

samples, then $r(\hat{\theta}) \leq \varepsilon$ with probability at least $1 - \delta$.

Praise and Critiques of Buzaglo et al.

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

- Student and teacher have the same depth.

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

- Student and teacher have the same depth.
- The parametrization we exploited is not necessary for generalization.

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

- Student and teacher have the same depth.
- The parametrization we exploited is not necessary for generalization.
- Posterior sampling may not be a good model.

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

- Student and teacher have the same depth.
- The parametrization we exploited is not necessary for generalization.
- Posterior sampling may not be a good model.
- Realizability assumption not met in practice.

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

- Student and teacher have the same depth.
- The parametrization we exploited is not necessary for generalization.
- Posterior sampling may not be a good model.
- Realizability assumption not met in practice.
- Quantization assumption doesn’t match practice, and is essential to the argument.

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

- Student and teacher have the same depth.
- The parametrization we exploited is not necessary for generalization.
- Posterior sampling may not be a good model.
- Realizability assumption not met in practice.
- Quantization assumption doesn’t match practice, and is essential to the argument.
- Predicts faster rate than seen empirically (in scaling laws).

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

- Student and teacher have the same depth.
- The parametrization we exploited is not necessary for generalization.
- Posterior sampling may not be a good model.
- Realizability assumption not met in practice.
- Quantization assumption doesn’t match practice, and is essential to the argument.
- Predicts faster rate than seen empirically (in scaling laws).
- Cannot explain performance as networks diverge in size.

Praise and Critiques of Buzaglo et al.

Nice observation that parametrization can induce “implicit bias”, even from “uniform” sampling.

On the other hand...

- Student and teacher have the same depth.
- The parametrization we exploited is not necessary for generalization.
- Posterior sampling may not be a good model.
- Realizability assumption not met in practice.
- Quantization assumption doesn’t match practice, and is essential to the argument.
- Predicts faster rate than seen empirically (in scaling laws).
- Cannot explain performance as networks diverge in size.

We’ll start with realizability (which will see us rethink posterior sampling, rates, etc.).

Primi: Gibbs Posterior

Primi: Gibbs Posterior

Non-realizable case means Θ^* empty.

Primi: Gibbs Posterior

Non-realizable case means Θ^* empty. Then possibly $\pi(\hat{\Theta}_0) = 0$ with positive probability, hence...

Primi: Gibbs Posterior

Non-realizable case means Θ^* empty. Then possibly $\pi(\hat{\Theta}_0) = 0$ with positive probability, hence... **no posterior**.

Primi: Gibbs Posterior

Non-realizable case means Θ^* empty. Then possibly $\pi(\hat{\Theta}_0) = 0$ with positive probability, hence... **no posterior**.

Gibbs posteriors allow us to model soft (stochastic) optimization and is well-defined in the non-realizable case.

Primi: Gibbs Posterior

Non-realizable case means Θ^* empty. Then possibly $\pi(\hat{\Theta}_0) = 0$ with positive probability, hence... **no posterior**.

Gibbs posteriors allow us to model soft (stochastic) optimization and is well-defined in the non-realizable case.

Defn. The *Gibbs posterior* for a prior π on Θ, inverse temperature $\beta > 0$, and empirical risk $\hat{r}_n$ under data S, is the distribution $\hat{\rho}$ on Θ dominated by π given by

$$\frac{\mathsf{d}\hat{\rho}}{\mathsf{d}\pi}(\theta) = \frac{1}{Z_\beta^\pi(S)} \exp\{-\beta\, \hat{r}_n(\theta)\} \propto \exp\{-\beta\, \hat{r}_n(\theta)\},$$

where $Z_\beta^\pi(S) := \int \exp\{-\beta\, \hat{r}_n(\theta)\}\pi(\mathsf{d}\theta)$ is the normalization constant.

Primi: Some defense for the Gibbs Posterior

Gibbs posterior:

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{1}{Z_\beta^\pi(S)} \exp\{-\beta\, \hat{r}_n(\theta)\}$$

$$\text{where } Z_\beta^\pi(S) := \int \exp\{-\beta\, \hat{r}_n(\theta)\}\pi(\mathrm{d}\theta).$$

Primi: Some defense for the Gibbs Posterior

Gibbs posterior:
$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{1}{Z_\beta^\pi(S)} \exp\{-\beta\, \hat{r}_n(\theta)\}$$
$$\text{where } Z_\beta^\pi(S) := \int \exp\{-\beta\, \hat{r}_n(\theta)\}\pi(\mathrm{d}\theta).$$

- **Generalizes hard posterior.** For *fixed* S, converges to (hard) posterior as inverse temperature $\beta \to \infty$.

Primi: Some defense for the Gibbs Posterior

Gibbs posterior:
$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{1}{Z_\beta^\pi(S)} \exp\{-\beta\, \hat{r}_n(\theta)\}$$
$$\text{where } Z_\beta^\pi(S) := \int \exp\{-\beta\, \hat{r}_n(\theta)\}\pi(\mathrm{d}\theta).$$

- **Generalizes hard posterior.** For *fixed* S, converges to (hard) posterior as inverse temperature $\beta \to \infty$.
- **Related to gradient flow + noise.** Assuming π is absolutely continuous with density p, the Gibbs posterior is the stationary distribution of Langevin dynamics with drift $\beta \hat{r}_n(\theta) + \log p(\theta)$.

Primi: Some defense for the Gibbs Posterior

Gibbs posterior:
$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{1}{Z_\beta^\pi(S)} \exp\{-\beta\, \hat{r}_n(\theta)\}$$
$$\text{where } Z_\beta^\pi(S) := \int \exp\{-\beta\, \hat{r}_n(\theta)\}\pi(\mathrm{d}\theta).$$

- **Generalizes hard posterior.** For *fixed* S, converges to (hard) posterior as inverse temperature $\beta \to \infty$.
- **Related to gradient flow + noise.** Assuming π is absolutely continuous with density p, the Gibbs posterior is the stationary distribution of Langevin dynamics with drift $\beta \hat{r}_n(\theta) + \log p(\theta)$.
- **Related to SGD + noise + decaying stepsize.** Gibbs posterior is also the limiting dynamics for Stochastic Gradient Langevin Dynamics (SGD + noise) for decaying step sizes.

Primi: Some defense for the Gibbs Posterior

Gibbs posterior: $$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{1}{Z_\beta^\pi(S)} \exp\{-\beta\, \hat{r}_n(\theta)\}$$

$$\text{where } Z_\beta^\pi(S) := \int \exp\{-\beta\, \hat{r}_n(\theta)\}\pi(\mathrm{d}\theta).$$

- **Generalizes hard posterior.** For *fixed* S, converges to (hard) posterior as inverse temperature $\beta \to \infty$.
- **Related to gradient flow + noise.** Assuming π is absolutely continuous with density p, the Gibbs posterior is the stationary distribution of Langevin dynamics with drift $\beta \hat{r}_n(\theta) + \log p(\theta)$.
- **Related to SGD + noise + decaying stepsize.** Gibbs posterior is also the limiting dynamics for Stochastic Gradient Langevin Dynamics (SGD + noise) for decaying step sizes.
- **Related to SGD + noise + fixed step size in large data limit.** The limiting OU process dynamics for SGLD and SGD in fixed step size regimes, where data, # of total iterations diverges. (These results don't cover high-dimensional case yet, though.)

Primi: Some defense for the Gibbs Posterior

Gibbs posterior:
$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \frac{1}{Z_\beta^\pi(S)} \exp\{-\beta\, \hat{r}_n(\theta)\}$$
$$\text{where } Z_\beta^\pi(S) := \int \exp\{-\beta\, \hat{r}_n(\theta)\}\pi(\mathrm{d}\theta).$$

- **Generalizes hard posterior.** For *fixed* S, converges to (hard) posterior as inverse temperature $\beta \to \infty$.
- **Related to gradient flow + noise.** Assuming π is absolutely continuous with density p, the Gibbs posterior is the stationary distribution of Langevin dynamics with drift $\beta\hat{r}_n(\theta) + \log p(\theta)$.
- **Related to SGD + noise + decaying stepsize.** Gibbs posterior is also the limiting dynamics for Stochastic Gradient Langevin Dynamics (SGD + noise) for decaying step sizes.
- **Related to SGD + noise + fixed step size in large data limit.** The limiting OU process dynamics for SGLD and SGD in fixed step size regimes, where data, # of total iterations diverges. (These results don't cover high-dimensional case yet, though.)
- **Gibbs posteriors optimize PAC-Bayes bounds.** The Gibbs posteriors above arises as the solution to variational problems: $\arg\min_\rho \rho(\hat{r}_n) + \beta^{-1} KL(\rho||\pi)$.
  That is, it minimizes certain PAC-Bayes bounds.

Primi: Single-sample PAC-Bayes Bound for Gibbs Posterior

Gibbs posterior: $\dfrac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta) = \dfrac{1}{Z_\beta^\pi(S)} \exp\{-\beta\, \hat{r}_n(\theta)\}$

where $Z_\beta^\pi(S) := \displaystyle\int \exp\{-\beta\, \hat{r}_n(\theta)\}\pi(\mathrm{d}\theta) = \pi(\exp\{-\beta \hat{r}_n\}).$

Corollary

Let $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature β, and $\hat{r}_n$, and let $\hat{\theta} \mid S \sim \hat{\rho}$. With probability at least $1-\delta$,

$$r(\hat{\theta}) \leq \Phi_{\lambda/n}^{-1}\Big[(1-\lambda^{-1}\beta)\hat{r}_n(\hat{\theta}) - \lambda^{-1}\log\big(Z_\beta^\pi(S)\big) + \lambda^{-1}\log\frac{1}{\delta}\Big].$$

If $\beta \geq \lambda$,

$$r(\hat{\theta}) \leq \Phi_{\lambda/n}^{-1}\Big[-\lambda^{-1}\log\big(Z_\beta^\pi(S)\big) + \lambda^{-1}\log\frac{1}{\delta}\Big].$$

Primi: Back to Teachers

Taking inverse temperature $\beta = \lambda$, we have a risk bound

Primi: Back to Teachers

Taking inverse temperature $\beta = \lambda$, we have a risk bound

$$r(\hat{\theta}) \leq \Phi_{\lambda/n}^{-1}\Big[-\lambda^{-1} \log \left(Z_\lambda^\pi(S)\right) + \lambda^{-1} \log \frac{1}{\delta}\Big].$$

Primi: Back to Teachers

Taking inverse temperature $\beta = \lambda$, we have a risk bound

$$r(\hat{\theta}) \leq \Phi^{-1}_{\lambda/n}\Big[-\lambda^{-1} \log \left(Z^{\pi}_{\lambda}(S)\right) + \lambda^{-1} \log \frac{1}{\delta}\Big].$$

We will see that this term is the **analogue of** $\lambda^{-1} \log \frac{1}{\pi(\Theta^*)}$.

Primi: Back to Teachers

Taking inverse temperature $\beta = \lambda$, we have a risk bound

$$r(\hat{\theta}) \leq \Phi^{-1}_{\lambda/n}\Big[-\lambda^{-1} \log \big(Z^{\pi}_{\lambda}(S)\big) + \lambda^{-1} \log \frac{1}{\delta}\Big].$$

We will see that this term is the **analogue of** $\lambda^{-1} \log \frac{1}{\pi(\Theta^*)}$.

Teacher lower bound on local entropy

Primi: Back to Teachers

Taking inverse temperature $\beta = \lambda$, we have a risk bound

$$r(\hat{\theta}) \leq \Phi^{-1}_{\lambda/n}\Big[-\lambda^{-1} \log \big(Z^{\pi}_{\lambda}(S)\big) + \lambda^{-1} \log \frac{1}{\delta}\Big].$$

We will see that this term is the **analogue of** $\lambda^{-1} \log \frac{1}{\pi(\Theta^*)}$.

Teacher lower bound on local entropy

Let $\theta^* \in \Theta$ be an (arbitrary!) teacher. Let $E_{\theta^*} = \{\theta \in \Theta : \theta^* \text{ and } \theta \text{ have same minimal model}\}$.

Primi: Back to Teachers

Taking inverse temperature $\beta = \lambda$, we have a risk bound

$$r(\hat{\theta}) \leq \Phi^{-1}_{\lambda/n}\Big[- \lambda^{-1} \log \big(Z^{\pi}_{\lambda}(S)\big) + \lambda^{-1} \log \frac{1}{\delta}\Big].$$

We will see that this term is the **analogue of** $\lambda^{-1} \log \frac{1}{\pi(\Theta^*)}$.

Teacher lower bound on local entropy

Let $\theta^* \in \Theta$ be an (arbitrary!) teacher. Let $E_{\theta^*} = \{\theta \in \Theta : \theta^* \text{ and } \theta \text{ have same minimal model}\}$.

$$Z^{\pi}_{\lambda}(S) = \int \exp\{-\lambda\, \hat{r}_n(\theta)\}\, \pi(\mathrm{d}\theta) \geq \exp\{-\lambda\, \hat{r}_n(\theta^*)\}\, \pi(E_{\theta^*}). \tag{1}$$

Primi: Back to Teachers

Taking inverse temperature $\beta = \lambda$, we have a risk bound

$$r(\hat{\theta}) \leq \Phi^{-1}_{\lambda/n}\Big[- \lambda^{-1} \log \big(Z^{\pi}_{\lambda}(S)\big) + \lambda^{-1} \log \frac{1}{\delta}\Big].$$

We will see that this term is the **analogue of** $\lambda^{-1} \log \frac{1}{\pi(\Theta^*)}$.

Teacher lower bound on local entropy

Let $\theta^* \in \Theta$ be an (arbitrary!) teacher. Let $E_{\theta^*} = \{\theta \in \Theta : \theta^* \text{ and } \theta \text{ have same minimal model}\}$.

$$Z^{\pi}_{\lambda}(S) = \int \exp\{-\lambda\, \hat{r}_n(\theta)\}\, \pi(\mathrm{d}\theta) \geq \exp\{-\lambda\, \hat{r}_n(\theta^*)\}\, \pi(E_{\theta^*}). \tag{1}$$

Thus, the local entropy obeys

$$-\lambda^{-1} \log \big(Z^{\pi}_{\lambda}(S)\big) \leq -\lambda^{-1} \log \big(\exp\{-\lambda\, \hat{r}_n(\theta^*)\}\, \pi(E_{\theta^*})\big) \tag{2}$$

$$= \hat{r}_n(\theta^*) + \lambda^{-1} \log \frac{1}{\pi(E_{\theta^*})} \tag{3}$$

Primi: Back to Teachers

Taking inverse temperature $\beta = \lambda$, we have a risk bound

$$r(\hat{\theta}) \leq \Phi^{-1}_{\lambda/n}\Big[-\lambda^{-1} \log \big(Z^{\pi}_{\lambda}(S)\big) + \lambda^{-1} \log \frac{1}{\delta}\Big].$$

We will see that this term is the **analogue of** $\lambda^{-1} \log \frac{1}{\pi(\Theta^*)}$.

Teacher lower bound on local entropy

Let $\theta^* \in \Theta$ be an (arbitrary!) teacher. Let $E_{\theta^*} = \{\theta \in \Theta : \theta^* \text{ and } \theta \text{ have same minimal model}\}$.

$$Z^{\pi}_{\lambda}(S) = \int \exp\{-\lambda\, \hat{r}_n(\theta)\}\, \pi(\mathrm{d}\theta) \geq \exp\{-\lambda\, \hat{r}_n(\theta^*)\}\, \pi(E_{\theta^*}). \tag{1}$$

Thus, the local entropy obeys

$$-\lambda^{-1} \log \big(Z^{\pi}_{\lambda}(S)\big) \leq -\lambda^{-1} \log \big(\exp\{-\lambda\, \hat{r}_n(\theta^*)\}\, \pi(E_{\theta^*})\big) \tag{2}$$

$$= \hat{r}_n(\theta^*) + \lambda^{-1} \log \frac{1}{\pi(E_{\theta^*})} \tag{3}$$

Indeed, the bound holds simultaneously for all $\theta^* \in \Theta^*$, and so we have

$$-\lambda^{-1} \log \big(Z^{\pi}_{\lambda}(S)\big) \leq \inf_{\theta^*} \Big\{ \hat{r}_n(\theta^*) + \lambda^{-1} \log \frac{1}{\pi(E_{\theta^*})} \Big\} \tag{4}$$

Primi: An oracle inequality

Primi: An oracle inequality

Key term is again $\log \frac{1}{\pi(E_{\theta^*})}$

Primi: An oracle inequality

Key term is again $\log \frac{1}{\pi(E_{\theta^*})} \leq m(\theta^*) \log Q$, by the same argument as Buzaglo et al.

Primi: An oracle inequality

Key term is again $\log \frac{1}{\pi(E_{\theta^*})} \leq m(\theta^*) \log Q$, by the same argument as Buzaglo et al.

Take $\theta^{**} \in \arg\min_{\theta \in \Theta^*} \left\{ r(\theta^*) + \lambda^{-1} m(\theta^*) \log Q \right\}$.

Primi: An oracle inequality

Key term is again $\log \frac{1}{\pi(E_{\theta^*})} \leq m(\theta^*) \log Q$, by the same argument as Buzaglo et al.

Take $\theta^{**} \in \arg\min_{\theta \in \Theta^*} \left\{ r(\theta^*) + \lambda^{-1} m(\theta^*) \log Q \right\}$.

With probability at least $1 - \delta$,

$$\begin{aligned} \inf_{\theta^* \in \Theta} \left\{ \hat{r}_n(\theta^*) + \lambda^{-1} m(\theta^*) \log Q \right\} &\leq r(\theta^{**}) + O(\sqrt{n^{-1} \log 1/\delta}) + \lambda^{-1} m(\theta^{**}) \log Q \\ &\leq \inf_{\theta^* \in \Theta} \left\{ r(\theta^*) + O(\sqrt{n^{-1} \log 1/\delta}) + \lambda^{-1} m(\theta^*) \log Q \right\} \end{aligned}$$

Primi: An oracle inequality

Key term is again $\log \frac{1}{\pi(E_{\theta^*})} \leq m(\theta^*) \log Q$, by the same argument as Buzaglo et al.

Take $\theta^{**} \in \arg\min_{\theta \in \Theta^*} \left\{ r(\theta^*) + \lambda^{-1} m(\theta^*) \log Q \right\}$.

With probability at least $1 - \delta$,

$$\begin{aligned} \inf_{\theta^* \in \Theta} \left\{ \hat{r}_n(\theta^*) + \lambda^{-1} m(\theta^*) \log Q \right\} &\leq r(\theta^{**}) + O(\sqrt{n^{-1} \log 1/\delta}) + \lambda^{-1} m(\theta^{**}) \log Q \\ &\leq \inf_{\theta^* \in \Theta} \left\{ r(\theta^*) + O(\sqrt{n^{-1} \log 1/\delta}) + \lambda^{-1} m(\theta^*) \log Q \right\} \end{aligned}$$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

Primi: An oracle inequality

Key term is again $\log \frac{1}{\pi(E_{\theta^*})} \leq m(\theta^*) \log Q$, by the same argument as Buzaglo et al.

Take $\theta^{**} \in \arg\min_{\theta \in \Theta^*} \left\{ r(\theta^*) + \lambda^{-1} m(\theta^*) \log Q \right\}$.

With probability at least $1 - \delta$,

$$\begin{aligned} \inf_{\theta^* \in \Theta} \left\{ \hat{r}_n(\theta^*) + \lambda^{-1} m(\theta^*) \log Q \right\} &\leq r(\theta^{**}) + O(\sqrt{n^{-1} \log 1/\delta}) + \lambda^{-1} m(\theta^{**}) \log Q \\ &\leq \inf_{\theta^* \in \Theta} \left\{ r(\theta^*) + O(\sqrt{n^{-1} \log 1/\delta}) + \lambda^{-1} m(\theta^*) \log Q \right\} \end{aligned}$$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

$$r(\hat{\theta}) \leq \inf_{\theta^* \in \Theta} O\left(r(\theta^*) + \sqrt{\frac{m(\theta^*) \log Q + \log 1/\delta}{n}} \right)$$

Can get a bound of the form $r + \frac{c}{n} + \sqrt{\frac{rc}{n}}$ too.

Primi: Toy MNIST Experiment

We validate our bounds on the MNIST dataset.

Experimental setup:

- 2-layer MLP with 3 hidden units (3167 parameters)
- Trained on MNIST dataset
- Obtained θ^* with 0.145 risk
- Fast rate bound yields 0.374 risk had we been able to quantize to 4-bits... but we haven't tried.

Primi: Takeaways

This bound balances teacher risk and complexity.

Key implications:

- Shifting to Gibbs posteriors allowed us to handle non-zero approximation error
- Provides insights into the trade-off between model complexity and fit

Secondi: Introducing $C(\varepsilon)$

Our oracle inequality doesn't tell us how risk behaves as we get more data.
We propose a new measure of data complexity based on teacher network size.

Secondi: Introducing $C(\varepsilon)$

Our oracle inequality doesn't tell us how risk behaves as we get more data.
We propose a new measure of data complexity based on teacher network size.

Defn. Let r^* is the Bayes error rate. Define

$$C(\varepsilon) = \inf_{\theta^* \in \Theta} m(\theta^*)$$
$$\text{s.t. } r(\theta^*) - r^* \leq \varepsilon$$

Secondi: Introducing $C(\varepsilon)$

Our oracle inequality doesn't tell us how risk behaves as we get more data.
We propose a new measure of data complexity based on teacher network size.

Defn. Let r^* is the Bayes error rate. Define

$$C(\varepsilon) = \inf_{\theta^* \in \Theta} m(\theta^*)$$
$$\text{s.t. } r(\theta^*) - r^* \leq \varepsilon$$

- Reflects the minimal network size needed for a given approximation error ε
- Monotonically increasing in ε
- Example of a rate–distortion function, up to some approximation.

Secondi: A bound with $C(\varepsilon)$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

Secondi: A bound with $C(\varepsilon)$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

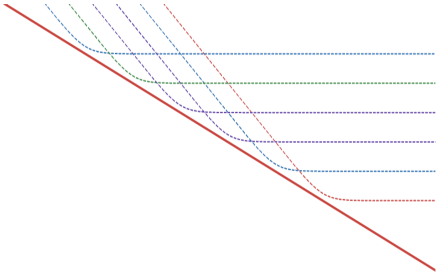
$$r(\hat{\theta}) \leq \inf_{\varepsilon} O\left(\varepsilon + \sqrt{\frac{\varepsilon\, C(\varepsilon) \log Q + \log 1/\delta}{n}}\right)$$

Secondi: A bound with $C(\varepsilon)$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

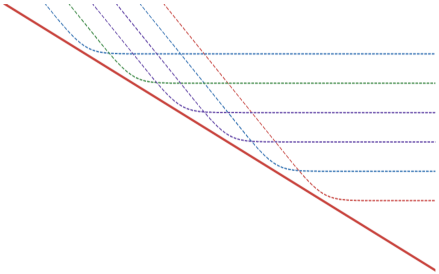
$$r(\hat{\theta}) \leq \inf_{\varepsilon} O\left(\varepsilon + \sqrt{\frac{\varepsilon\, C(\varepsilon) \log Q + \log 1/\delta}{n}}\right)$$



Secondi: Interpreting $C(\varepsilon)$

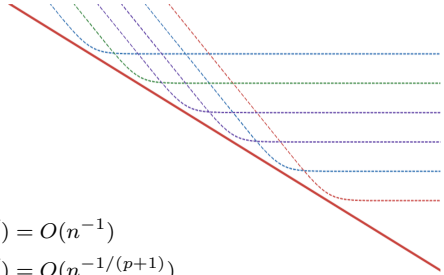
Secondi: Interpreting $C(\varepsilon)$

$$r(\hat{\theta}) \leq \inf_{\varepsilon} O\left(\varepsilon + \sqrt{\frac{\varepsilon\, C(\varepsilon) \log Q + \log 1/\delta}{n}}\right)$$



Secondi: Interpreting $C(\varepsilon)$

$$r(\hat{\theta}) \leq \inf_{\varepsilon} O\left(\varepsilon + \sqrt{\frac{\varepsilon\, C(\varepsilon) \log Q + \log 1/\delta}{n}}\right)$$

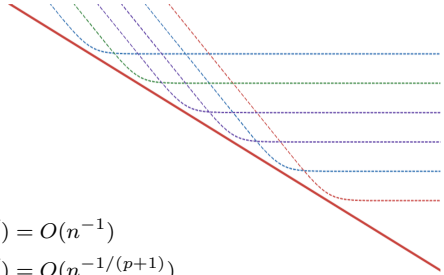


The "complexity" function $C(\varepsilon)$ dictates risk rate:

if $C(\varepsilon) = O(\text{polylog}(\varepsilon^{-1}))$	then $r(\hat{\theta}) = O(n^{-1})$
if $C(\varepsilon) = O(\varepsilon^{-p})$	then $r(\hat{\theta}) = O(n^{-1/(p+1)})$
if $C(\varepsilon) = O(\exp(\text{poly}(\varepsilon^{-1})))$	then $r(\hat{\theta}) = O(\log^{-1} n)$.

Secondi: Interpreting $C(\varepsilon)$

$$r(\hat{\theta}) \leq \inf_{\varepsilon} O\left(\varepsilon + \sqrt{\frac{\varepsilon\, C(\varepsilon) \log Q + \log 1/\delta}{n}}\right)$$



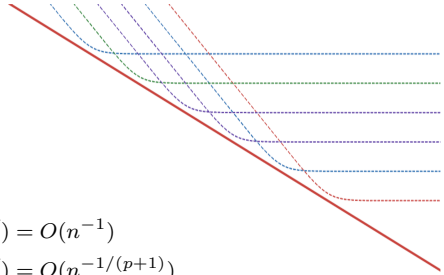
The "complexity" function $C(\varepsilon)$ dictates risk rate:

if $C(\varepsilon) = O(\text{polylog}(\varepsilon^{-1}))$	then $r(\hat{\theta}) = O(n^{-1})$
if $C(\varepsilon) = O(\varepsilon^{-p})$	then $r(\hat{\theta}) = O(n^{-1/(p+1)})$
if $C(\varepsilon) = O(\exp(\text{poly}(\varepsilon^{-1})))$	then $r(\hat{\theta}) = O(\log^{-1} n)$.

Consequence: If scaling law for n data and size $m(n)$ model says risk decays *slower* than $O(n^{-1/(p+1)})$, then...

Secondi: Interpreting $C(\varepsilon)$

$$r(\hat{\theta}) \leq \inf_{\varepsilon} O\left(\varepsilon + \sqrt{\frac{\varepsilon\, C(\varepsilon) \log Q + \log 1/\delta}{n}}\right)$$

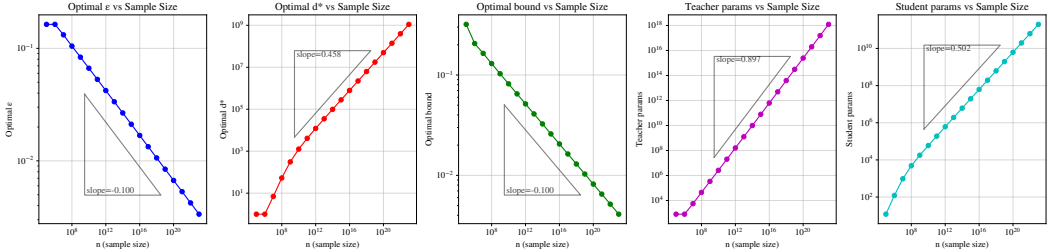


The "complexity" function $C(\varepsilon)$ dictates risk rate:

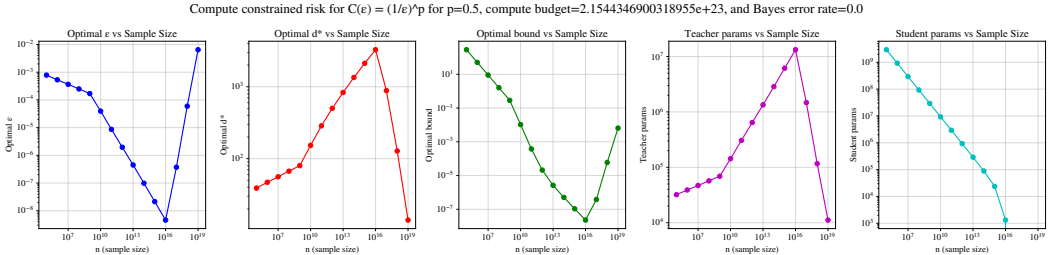
if $C(\varepsilon) = O(\text{polylog}(\varepsilon^{-1}))$	then $r(\hat{\theta}) = O(n^{-1})$
if $C(\varepsilon) = O(\varepsilon^{-p})$	then $r(\hat{\theta}) = O(n^{-1/(p+1)})$
if $C(\varepsilon) = O(\exp(\text{poly}(\varepsilon^{-1})))$	then $r(\hat{\theta}) = O(\log^{-1} n)$.

Consequence: If scaling law for n data and size $m(n)$ model says risk decays *slower* than $O(n^{-1/(p+1)})$, then... $C(\varepsilon) \notin O(\varepsilon^{-p})$.

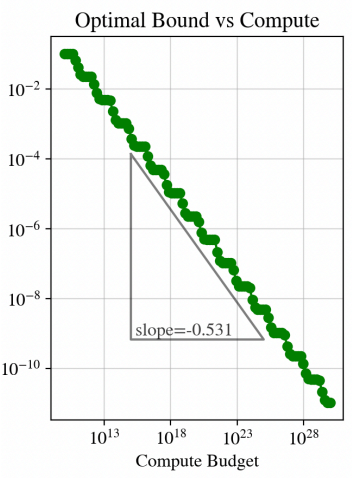
Scaling laws for $C(\varepsilon) = \varepsilon^{-9}$.



Risk bounds for fixed compute, for $C(\varepsilon) = \varepsilon^{-9}$.



Compute optimal scaling laws for $C(\varepsilon) = \varepsilon^{-9}$.



Dolce: Stochastic teachers

Donsker–Varadhan offers a more sophisticated lower bound on local entropy.

Dolce: Stochastic teachers

Donsker–Varadhan offers a more sophisticated lower bound on local entropy.

$$-\lambda^{-1} \log \left(Z_{\lambda}^{\pi}(S) \right) = \inf_{\rho} \rho(\hat{r}_n) + \lambda^{-1} KL(\rho||\pi) \tag{5}$$

Dolce: Stochastic teachers

Donsker–Varadhan offers a more sophisticated lower bound on local entropy.

$$-\lambda^{-1} \log \left(Z_\lambda^\pi(S) \right) = \inf_\rho \rho(\hat{r}_n) + \lambda^{-1} KL(\rho||\pi) \tag{5}$$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

Dolce: Stochastic teachers

Donsker–Varadhan offers a more sophisticated lower bound on local entropy.

$$-\lambda^{-1} \log \left(Z_\lambda^\pi(S) \right) = \inf_\rho \rho(\hat{r}_n) + \lambda^{-1} KL(\rho||\pi) \tag{5}$$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

$$r(\hat{\theta}) \leq \inf_{\rho^* \in \Delta(\Theta)} O\left(\rho(r) + \sqrt{\frac{KL(\rho||\pi) + \log 1/\delta}{n}} \right)$$

Dolce: Stochastic teachers

Donsker–Varadhan offers a more sophisticated lower bound on local entropy.

$$-\lambda^{-1} \log \left(Z_\lambda^\pi(S) \right) = \inf_\rho \rho(\hat{r}_n) + \lambda^{-1} KL(\rho||\pi) \tag{5}$$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

$$r(\hat{\theta}) \leq \inf_{\rho^* \in \Delta(\Theta)} O\left(\rho(r) + \sqrt{\frac{KL(\rho||\pi) + \log 1/\delta}{n}} \right)$$

- Every PAC-Bayes bound (regardless of the predictor) offers a risk bound for the Gibbs posterior-sampled predictor.

Dolce: Stochastic teachers

Donsker–Varadhan offers a more sophisticated lower bound on local entropy.

$$-\lambda^{-1} \log \left(Z_\lambda^\pi(S) \right) = \inf_\rho \rho(\hat{r}_n) + \lambda^{-1} KL(\rho||\pi) \tag{5}$$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

$$r(\hat{\theta}) \leq \inf_{\rho^* \in \Delta(\Theta)} O\left(\rho(r) + \sqrt{\frac{KL(\rho||\pi) + \log 1/\delta}{n}} \right)$$

- Every PAC-Bayes bound (regardless of the predictor) offers a risk bound for the Gibbs posterior-sampled predictor.
- Offers an explicit connection with coding. Cf. $C(\varepsilon)$ as minimal teacher size.

Dolce: Stochastic teachers

Donsker–Varadhan offers a more sophisticated lower bound on local entropy.

$$-\lambda^{-1} \log \left(Z_\lambda^\pi(S) \right) = \inf_\rho \rho(\hat{r}_n) + \lambda^{-1} KL(\rho||\pi) \tag{5}$$

Theorem

There exists $\lambda > 0$ such that, letting $\hat{\rho}$ be the Gibbs posterior for π, inverse temperature λ, and $\hat{r}_n$, and letting $\hat{\theta} \mid S \sim \hat{\rho}$, with probability at least $1 - \delta$,

$$r(\hat{\theta}) \leq \inf_{\rho^* \in \Delta(\Theta)} O\left(\rho(r) + \sqrt{\frac{KL(\rho||\pi) + \log 1/\delta}{n}} \right)$$

- Every PAC-Bayes bound (regardless of the predictor) offers a risk bound for the Gibbs posterior-sampled predictor.
- Offers an explicit connection with coding. Cf. $C(\varepsilon)$ as minimal teacher size.
- For analysis, we could potentially look to Hessians/flatness (Yang, Mao, Chaudhari 2022). Cf. Dziugaite & Roy 2017.

Key Insights

Our work bridges theory and practice in deep learning generalization.

Main contributions:

- Theoretical explanation for generalization in (mildly) overparameterized models
- Novel measure of data complexity via teacher network size
- Insights into the role of data complexity in learning
- Numerical scaling laws offer us evidence of complexity via excess risk upper bounds

Limitations

Our work has some limitations that offer opportunities for future research.

Current limitations:

- Naive approach to lower bounding probability of interpolation
- Bounds tightest when student size doesn't exceed teacher size

Conclusion

We've introduced a novel perspective on data complexity in deep learning.

Key contributions:

- Novel measure of data complexity based on teacher network size
- Non-vacuous generalization bounds for neural networks
- Connection between theoretical bounds and empirical scaling laws