SLT High 1
Logic of Phase Transitions

The Primer

Key Ideas

- **Thermodynamics** says the world is organised by *phases* and *phase transitions*.
- **Geometrically** that means *singularities* and *unfoldings*.
- **Singular Learning Theory** connects both to *learning machines*.
- If neural network structure forms in discrete phase transitions like those of developmental biology, this should be a powerful language for **interpretability**.

The Primer

Key Ideas

- The **Free Energy Formula** is the main theoretical result of SLT. It expresses free energy in terms of energy and the learning coefficient

$$F_n = nL_n(w_0) + \lambda \log n$$

- The power of thermodynamics is that you can derive from a simple **fundamental relation**, using nothing more than first year calculus and algebra, nontrivial predictions about physical systems.
- **Aim of the Primer:** explain the Free Energy Formula, and demonstrate in an example of interest in AI alignment (Toy Models of Superposition) how to derive nontrivial statements from it.

References

- **The Gray Book,** S. Watanabe "Algebraic Geometry and Statistical Learning Theory", 2009.
- **The Green Book**, S. Watanabe "Mathematical Theory of Bayesian Statistics", 2018.
- **The WBIC paper**, S. Watanabe "A Widely Applicable Bayesian Information Criterion" JMLR 2013.
- **The renormalizability paper**, S. Watanabe "Asymptotic learning curve and renormalizable condition in statistical learning theory" Journal of Physics 2010.
- S. Watanabe "**Cross Validation, Information Criterion and Phase Transition**", talk 2023 (earlier talk Phase Transition and Prior Effect).

SLT High 1

Logic of Phase Transitions

- The Free Energy Formula
- Model selection as Coarse-Graining
- Internal model selection
- Thermodynamics
- Singular Learning Process

Setup

- Samples $X_1, \ldots, X_n$ are independently subject to a true distribution $q(x)$. We denote by $p(x \mid w)$ our model and $\varphi(w)$ our prior, on parameter space W.
- The log loss is $L_n(w) = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \log p(X_i \mid w)$
- The (Bayes) free energy is defined to be

$$
\begin{aligned}
F_n &= -\log \int \prod_{i=1}^{n} p(X_i \mid w)\varphi(w)dw \\
&= -\log \int \exp(-nL_n(w))\varphi(w)dw
\end{aligned}
$$

Setup

For Neural Networks

- The true distribution is $q(x, y) = q(y \mid x)q(x)$ with inputs $x \in \mathbb{R}^m$ and outputs $y \in \mathbb{R}^n$. We denote by $p(x, y \mid w) = p(y \mid x, w)q(x)$ our model and $\varphi(w)$ our prior, on parameter space W. Suppose given samples $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- The model is given by $p(y \mid x, w) = \dfrac{1}{(2\pi)^{n/2}} \exp\left(-\dfrac{1}{2} \left\| y - f(x, w) \right\|^2\right)$ where $f(x, w)$ is a neural network with weights w.
- In this case the log loss is the mean squared error (up to some constants).

Setup

For Neural Networks

$$
\begin{aligned}
L_n(w) &= -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i, Y_i \mid w) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \left\| Y_i - f(X_i, w) \right\|^2\right) q(X_i)\right] \\
&= \frac{1}{2n}\sum_{i=1}^{n} \left\| Y_i - f(X_i, w) \right\|^2 - \frac{1}{n}\sum_{i=1}^{n} \log q(X_i) + \text{const.}
\end{aligned}
$$

Mean squared error, i.e. "loss"

Empirical entropy of $q(x)$

Setup

For Neural Networks

$$L_n(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left\| Y_i - f(X_i, w)\right\|^2 - \frac{1}{n}\sum_{i=1}^{n}\log q(X_i) + \text{const.}$$

$$\begin{aligned} F_n &= -\log\int\prod_{i=1}^{n} p(X_i \mid w)\varphi(w)dw \\ &= -\log\int\exp(-nL_n(w))\varphi(w)dw \\ &= -\log Z_n \end{aligned}$$

Partition function / model evidence

$$Z_n = \int\exp(-nL_n(w))\varphi(w)dw$$

Bayesian posterior

$$p(w \mid D_n) = \frac{1}{Z_n}\exp(-nL_n(w))\varphi(w)$$

Free Energy Formula

Precise Statement

- Assume *relative finite variance* [**Green**, §3.1] in addition to the fundamental conditions of [**Gray**] (excepting realisability) and that there is a point w_0 minimising L in the interior of W.
- **Theorem** (Watanabe): We have by [**Green**, §6.3], see also [**WBIC, Renormalizability**]:

$$F_n = nL_n(w_0) + \lambda \log n - (m-1)\log\log n + F_n^R + o_p(1)$$

- Here $\lambda \in \mathbb{Q}_{>0}$ is called the *learning coefficient*, $m \in \mathbb{N}$ is the *multiplicity* and F_n^R is a random variable which converges to a random variable in law.

Free Energy Formula

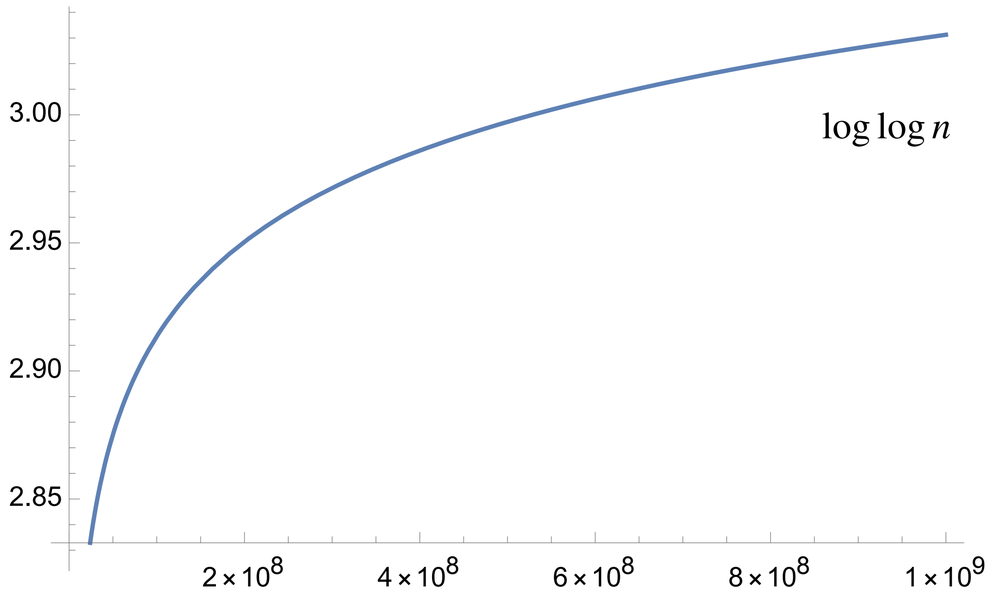
Friendly Statement

$$F_n = nL_n(w_0) + \lambda \log n + O_p(\log \log n)$$

Free Energy Formula

Really Friendly Statement

$$F_n = nL_n(w_0) + \lambda \log n + const$$



Model Selection

$q(x), D_n$

Model 1: p_1, W_1

$$F_{p_1}(D_n, W_1) = -\log \int_{W_1} \prod_{i=1}^{n} p_1(X_i|w)\,\phi_1(w)\,dw$$

Model 2: p_2, W_2

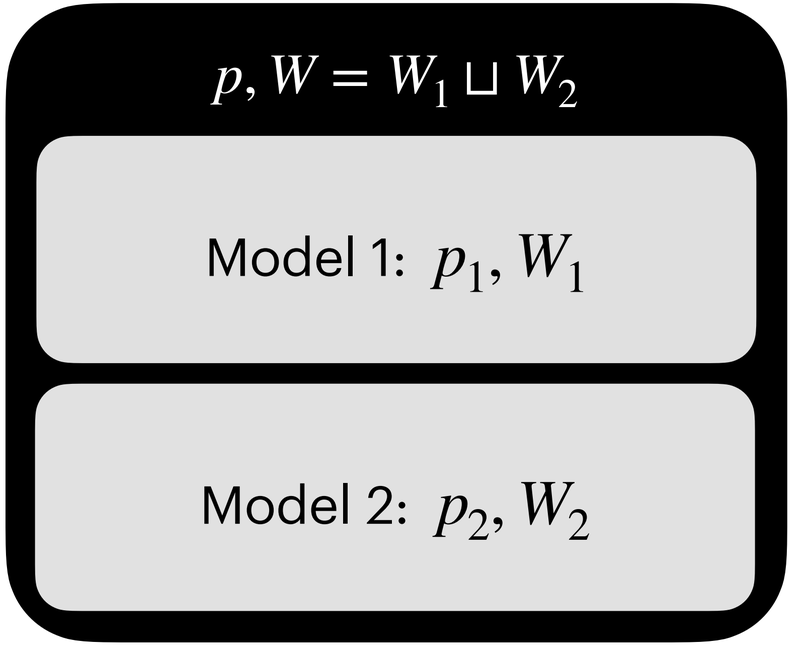
$$F_{p_2}(D_n, W_2) = -\log \int_{W_2} \prod_{i=1}^{n} p_2(X_i|w)\,\phi_2(w)\,dw$$

Dogma: prefer the model with lowest free energy.

Model Selection

Models As States

$q(x), D_n$



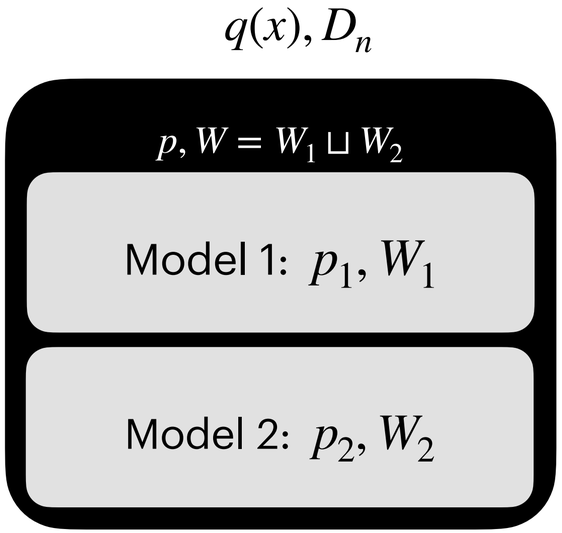
$$F_n(W_1) = -\log \int_{W_1} \prod_{i=1}^{n} p_1(X_i \mid w)\varphi(w)dw$$

$$F_n(W_2) = -\log \int_{W_2} \prod_{i=1}^{n} p_2(X_i \mid w)\varphi(w)dw$$

$$F_n(W) = -\log \int_{W} \prod_{i=1}^{n} p_1(X_i \mid w)\varphi(w)dw = -\log\left[e^{-F_n(W_1)} + e^{-F_n(W_2)}\right]$$

Model Selection

Models As States



$$F_n(W_\alpha) = -\log \int_{W_\alpha} \prod_{i=1}^{n} p_1(X_i \mid w)\varphi(w)dw$$

$$F_n(W) = -\log \int_{W} \prod_{i=1}^{n} p_1(X_i \mid w)\varphi(w)dw = -\log\Big[\sum_{\alpha} e^{-F_n(W_\alpha)}\Big]$$

Meta-prior

$$p(W_\alpha \mid D_n) = \frac{p(D_n \mid W_\alpha)p(W_\alpha)}{p(D_n)} = \frac{e^{-F_n(W_\alpha)}\Phi_\alpha}{\sum_\beta e^{-F_n(W_\beta)}\Phi_\beta}$$

Model evidence (in context)

Model Selection As Coarse-Graining

	Fine-grained	**Coarse-grained**
Microstate	w	α
Microscopic energy	$nL_n(w) - \log \varphi(w)$	$F_n(W_\alpha) - \log \Phi_\alpha$
Boltzmann distribution	$e^{-nL_n(w)}\varphi(w)$	$e^{-F_n(W_\alpha)}\Phi_\alpha$
Partition function	$\int_{W_\alpha} e^{-nL_n(w)}\varphi(w)dw = e^{-F_n(W_\alpha)}$	$\mathfrak{Z}_n = \sum_\beta e^{-F_n(W_\beta)}\Phi_\beta$

$$p(W_\alpha \mid D_n) = \frac{p(D_n \mid W_\alpha)p(W_\alpha)}{p(D_n)} = \frac{e^{-F_n(W_\alpha)}\Phi_\alpha}{\sum_\beta e^{-F_n(W_\beta)}\Phi_\beta}$$

Model Selection As Coarse-Graining

- Bayes Rule says that model selection is governed by the **coarse-grained Boltzmann distribution** with the free energy $F_n(W_\alpha)$ as the new microscopic Hamiltonian for the model index α (the emergent coordinate).
- For large differences in free energy, it's similar to just selecting the model α with the lowest free energy.
- This Boltzmann distribution is parametrised by n, and it's interesting to think about what happens as this parameter is varied.

Model Selection As Coarse-Graining

	Fine-grained	**Coarse-grained**
Microstate	w	α
Microscopic energy	$nL_n(w) - \log \varphi(w)$	$F_n(W_\alpha) - \log \Phi_\alpha$
Boltzmann distribution	$e^{-nL_n(w)}\varphi(w)$	$e^{-F_n(W_\alpha)}\Phi_\alpha$
Partition function	$\int_{W_\alpha} e^{-nL_n(w)}\varphi(w)dw = e^{-F_n(W_\alpha)}$	$\mathfrak{Z}_n = \sum_\beta e^{-F_n(W_\beta)}\Phi_\beta$

$$p(W_\alpha \mid D_n) = \frac{p(D_n \mid W_\alpha)p(W_\alpha)}{p(D_n)} = \frac{e^{-F_n(W_\alpha)}\Phi_\alpha}{\sum_\beta e^{-F_n(W_\beta)}\Phi_\beta}$$

Model Selection As Coarse-Graining

$$
\begin{aligned}
\mathfrak{F}_n &:= -\log \mathfrak{Z}_n \\
&= -\log\Big[\sum_{\beta} \exp(-F_n(W_\beta))\Big] \\
&\approx -\max_{\beta}(-F_n(W_\beta)) \\
&= \min_{\beta} F_n(W_\beta)
\end{aligned}
$$

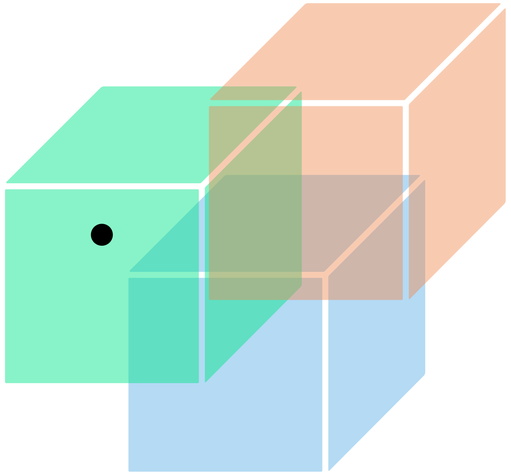
- The coarse-grained free energy $\mathfrak{F}_n$ approximately minimises the free energy over the available models β.

Internal Model Selection

- Model selection is usually thought of something that statisticians do.
- However, the story above about coarse-graining is mostly interesting because it happens **automatically** in Bayesian learning, **internally** to a single model.
- Given a model (p, q, φ) with parameter space W we refer to the emergent submodels W_α, between which this internal model selection chooses, as *phases*. A change in n or another hyperparameter which leads to a different choice is called a *phase transition*.

Internal Model Selection

- Following Watanabe's talk on phase transitions, let $\{W_\alpha\}_\alpha$ be a finite collection of compact sets that cover W, where each W_α is semi-analytic, has non-empty interior, contains in its interior a point w_α^* that minimises $L(w)$ on W_α and has relative finite variance. We assume that every point $w \in W$ lies in the interior of some W_α.



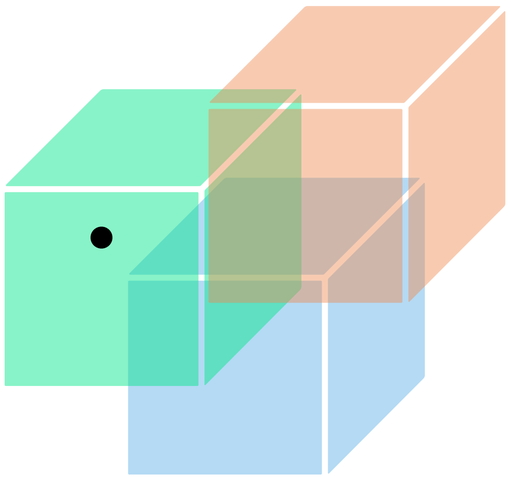
- Let $\{U_{\alpha,\gamma}\}_{\alpha,\gamma}$ be a finite open cover of W with the property that $U_{\alpha,\gamma} \subseteq W_\alpha$ for all γ and let $\rho_{\alpha,\gamma}$ be a partition of unity subordinate to this cover.
- Set $\rho_\alpha = \sum_\gamma \rho_{\alpha,\gamma}$ and $\varphi_\alpha(w) = \rho_\alpha(w)\varphi(w)$.

Internal Model Selection

$$F_n = -\log \int_W e^{-nL_n(w)} \varphi(w) dw$$

$$= -\log \sum_\alpha \int_{W_\alpha} e^{-nL_n(w)} \varphi_\alpha(w) dw$$

$$= -\log \sum_\alpha e^{-F_n(W_\alpha)}$$



- Here $F_n(W_\alpha) = -\log \int_{W_\alpha} exp(-nL_n(w))\varphi_\alpha(w)dw$ is the free energy of the submodel with parameter space W_α, prior φ_α, and the same model p, truth q as the original.

Internal Model Selection

- We can apply the Free Energy Formula to the model $(p, q, \varphi_\alpha, W_\alpha)$

$$F_n(W_\alpha) = nL_n(w_\alpha^*) + \lambda_\alpha \log n + c_\alpha$$

- Then

$$F_n = -\log \sum_\alpha e^{-F_n(W_\alpha)} \approx \min_\alpha F_n(W_\alpha)$$

$$\approx \min_\alpha \left[nL_n(w_\alpha^*) + \lambda_\alpha \log n + c_\alpha \right]$$

- The Bayesian posterior **selects** phases on the basis of competition between *energy, complexity* and subleading terms (which include prior effects). When the index α changes as a function of n or hyperparameters, we say that there has been a *phase transition* in the Bayesian posterior.

Thermodynamics

- Now we take the **Free Energy Formula** and the principle of **Internal Model Selection** and do thermodynamics, that is, we deduce several interesting facts about learning machines from elementary manipulations of the formula

$$F_n = -\log \sum_{\alpha} e^{-F_n(W_\alpha)} \approx \min_{\alpha} F_n(W_\alpha)$$
$$\approx \min_{\alpha} \left[nL_n(w_\alpha^*) + \lambda_\alpha \log n + c_\alpha \right]$$

- We make two additional simplifying assumptions: replacing $L_n(w_\alpha^*)$ by the deterministic $L(w_\alpha^*)$ and assuming that $c_\alpha = 0$.

Thermodynamics

- Now we take the **Free Energy Formula** and the principle of **Internal Model Selection** and do thermodynamics, that is, we deduce several interesting facts about learning machines from elementary manipulations of the formula

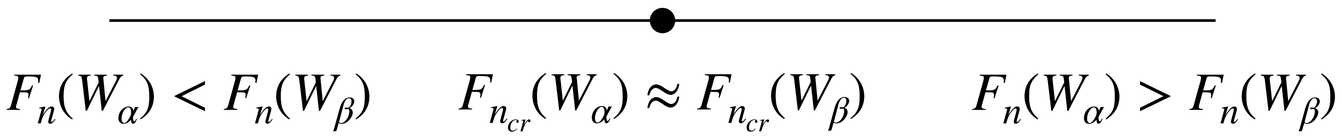
$$F_n \approx \min_{\alpha} F_n(W_\alpha) \approx \min_{\alpha} \left[nE_\alpha + \lambda_\alpha \log n \right]$$

- Here $E_\alpha = L(w_\alpha^*)$ is the *energy* of the phase α and λ_α is the *learning coefficient* which is a measure of *complexity*.
- In the following indices $\alpha, \beta, \gamma, \ldots$ stand for phases.

Thermodynamics

$$F_n \approx \min_\alpha F_n(W_\alpha) \approx \min_\alpha \left[nE_\alpha + \lambda_\alpha \log n \right]$$

- If a phase α is dominated by a phase β both with respect to energy $E_\alpha > E_\beta$ and learning coefficient $\lambda_\alpha > \lambda_\beta$ then $F_n(W_\alpha) > F_n(W_\beta)$ but there is **no phase transition** because this is true for all n.
- For there to be a phase transition in n between phases $\alpha \longrightarrow \beta$ we need both a *critical dataset size* $n = n_{cr}$ and for this transition to not be "screened" by others:



Thermodynamics

$$F_n \approx \min_\alpha F_n(W_\alpha) \approx \min_\alpha \left[nE_\alpha + \lambda_\alpha \log n\right]$$

- Assume without loss of generality that $E_\alpha > E_\beta$ and $\lambda_\alpha < \lambda_\beta$. Then

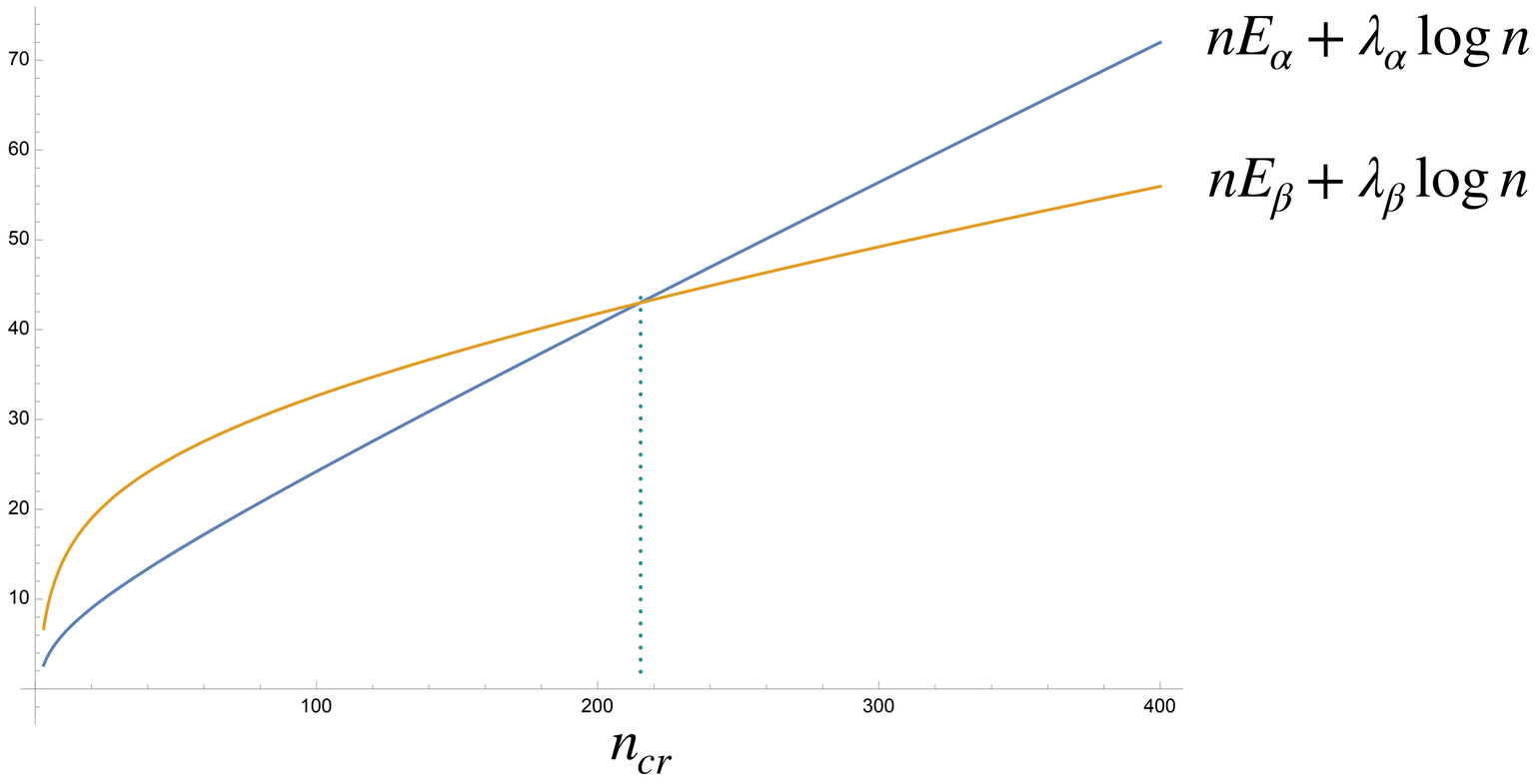
$$\begin{aligned} F_n(W_\alpha) = F_n(W_\beta) \quad &\iff \quad nE_\alpha + \lambda_\alpha \log n = nE_\beta + \lambda_\beta \log n \\ &\iff \quad n(E_\alpha - E_\beta) = -\log n(\lambda_\alpha - \lambda_\beta) \\ &\iff \quad \frac{n}{\log n} = -\frac{\lambda_\beta - \lambda_\alpha}{E_\beta - E_\alpha} \end{aligned}$$

- The function $n/\log n$ is positive and increasing for $n > e$ so this has a unique solution, which is the critical dataset size n_{cr}.

Thermodynamics

$$F_n \approx \min_{\alpha} F_n(W_\alpha) \approx \min_{\alpha} \left[nE_\alpha + \lambda_\alpha \log n \right]$$

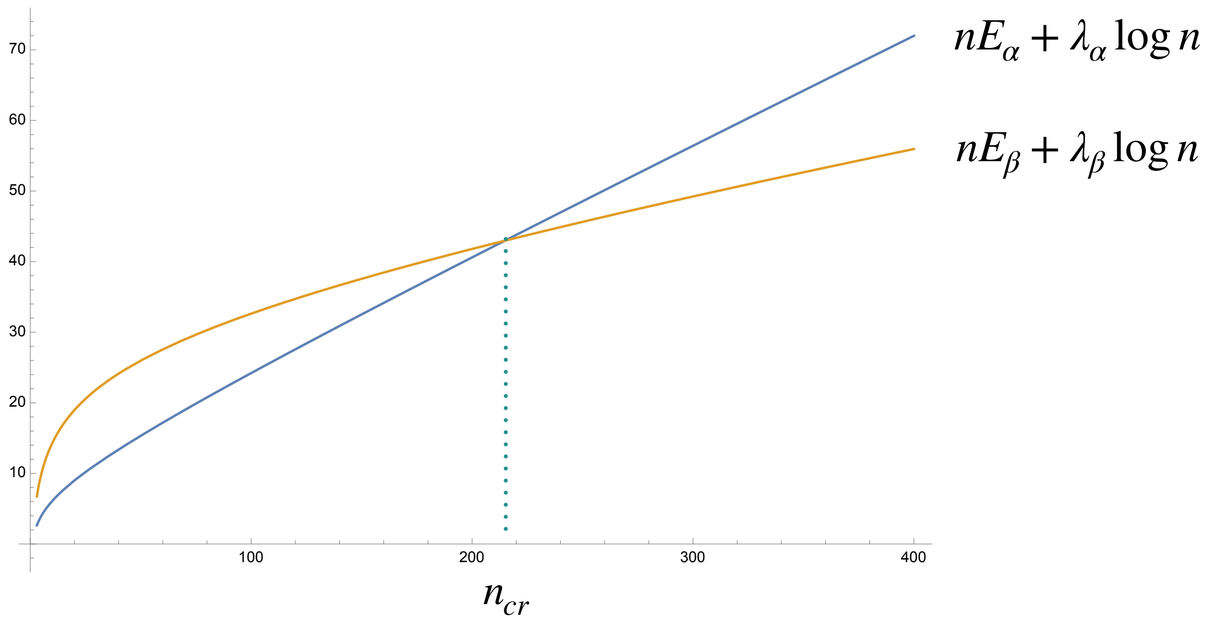
- If $E_\alpha > E_\beta$ and $\lambda_\alpha < \lambda_\beta$ then there is a (candidate) transition $\alpha \longrightarrow \beta$

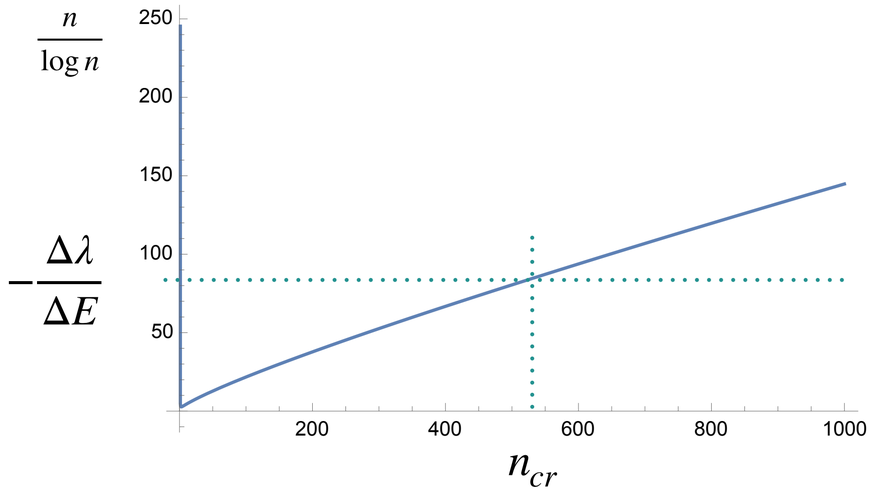


Thermodynamics

$$F_n \approx \min_{\alpha} F_n(W_\alpha) \approx \min_{\alpha} \left[nE_\alpha + \lambda_\alpha \log n \right]$$

- If $E_\alpha > E_\beta$ and $\lambda_\alpha < \lambda_\beta$ then there is a (candidate) transition $\alpha \longrightarrow \beta$





$$n_{cr} = \mathcal{N}\left(-\frac{\Delta\lambda}{\Delta E}\right) \qquad \mathcal{N} \text{ is inverse to } \frac{n}{\log n}$$

Thermodynamics

$$F_n \approx \min_{\alpha} F_n(W_\alpha) \approx \min_{\alpha} \left[nE_\alpha + \lambda_\alpha \log n \right]$$

- **Observation 1.** Assuming that $E_\alpha > E_\beta$ and $\lambda_\alpha < \lambda_\beta$ there is a (candidate) phase transition in the Bayesian posterior $\alpha \longrightarrow \beta$ at $n = n_{cr} = \mathcal{N}\left(-\frac{\Delta\lambda}{\Delta E}\right)$. We call this the *critical dataset size* for the transition.

Transition?	$E_\alpha > E_\beta$	$E_\alpha < E_\beta$
$\lambda_\alpha < \lambda_\beta$	$\alpha \longrightarrow \beta$	No
$\lambda_\alpha > \lambda_\beta$	No	$\beta \longrightarrow \alpha$

Thermodynamics

$$F_n \approx \min_\alpha F_n(W_\alpha) \approx \min_\alpha \left[nE_\alpha + \lambda_\alpha \log n \right]$$

- **Observation 1.** Assuming that $E_\alpha > E_\beta$ and $\lambda_\alpha < \lambda_\beta$ there is a (candidate) phase transition in the Bayesian posterior $\alpha \longrightarrow \beta$ at $n = n_{cr} = \mathcal{N}\left(-\frac{\Delta\lambda}{\Delta E}\right)$. We call this the *critical dataset size* for the transition.

- **Observation 2.** Phase transitions in n that change the energy must *decrease* the energy and *increase* the learning coefficient.

"The learning process produces *more accurate* models that are *more complex,* sacrificing extra bits in the model description for fewer errors"

Thermodynamics

$$F_n \approx \min_\alpha F_n(W_\alpha) \approx \min_\alpha \left[nE_\alpha + \lambda_\alpha \log n + c_\alpha\right]$$

- If $E_\alpha = E_\beta$ and the subleading terms are equal $c_\alpha = c_\beta$ then there is no transition

$$F_n(W_\alpha) = F_n(W_\beta) \iff 0 = \log n(\lambda_\beta - \lambda_\alpha) \iff \lambda_\beta = \lambda_\alpha$$

- Assume that $E_\alpha = E_\beta$ and $c_\alpha > c_\beta$. Then

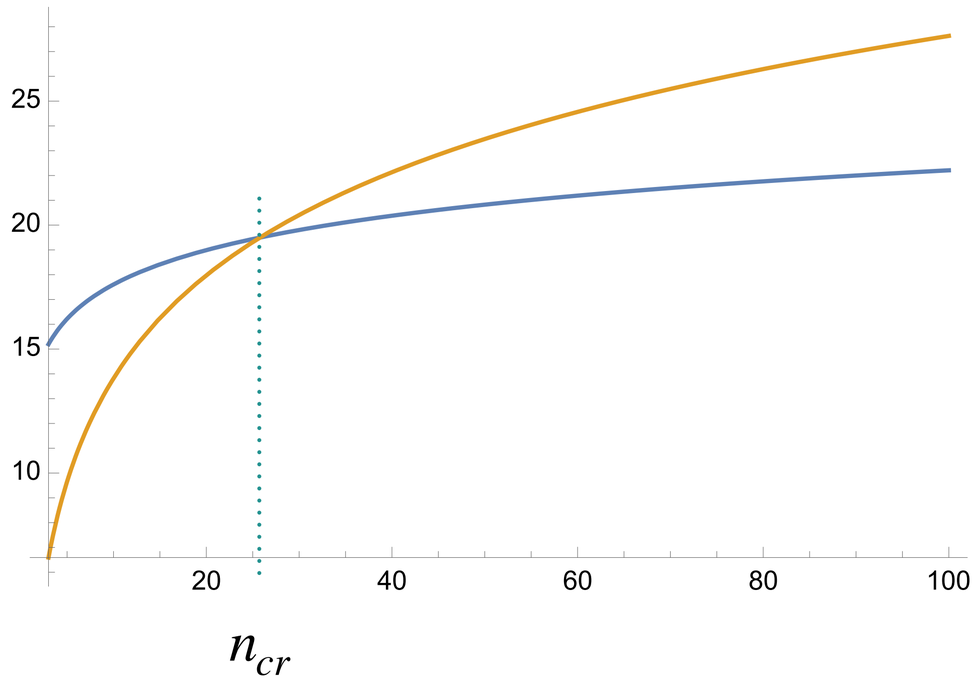
$$\begin{aligned} F_n(W_\alpha) = F_n(W_\beta) \quad &\iff \quad nE_\alpha + \lambda_\alpha \log n + c_\alpha = nE_\beta + \lambda_\beta \log n + c_\beta \\ &\iff \quad c_\alpha - c_\beta = \log n(\lambda_\beta - \lambda_\alpha) \\ &\iff \quad -\frac{c_\beta - c_\alpha}{\lambda_\beta - \lambda_\alpha} = \log n \end{aligned}$$

- This has a solution if and only if $\lambda_\beta > \lambda_\alpha$ in which case it is $n_{cr} = \exp\left(-\frac{\Delta c}{\Delta \lambda}\right)$.

Thermodynamics

$$F_n \approx \min_\alpha F_n(W_\alpha) \approx \min_\alpha \left[nE_\alpha + \lambda_\alpha \log n + c_\alpha\right]$$

- If $E_\alpha = E_\beta$ and $c_\alpha > c_\beta$ and $\lambda_\beta > \lambda_\alpha$ there is a (candidate) transition $\beta \longrightarrow \alpha$.



$\lambda_\beta \log n + c_\beta$

$\lambda_\alpha \log n + c_\alpha$

$$n_{cr} = \exp\left(-\frac{\Delta c}{\Delta \lambda}\right)$$

Thermodynamics

$$F_n \approx \min_\alpha F_n(W_\alpha) \approx \min_\alpha \left[nE_\alpha + \lambda_\alpha \log n \right]$$

- **Law 1.** Assuming that $E_\alpha > E_\beta$ and $\lambda_\alpha < \lambda_\beta$ there is a (candidate) phase transition in the Bayesian posterior $\alpha \longrightarrow \beta$ at $n = n_{cr} = \mathcal{N}\left(-\frac{\Delta\lambda}{\Delta E}\right)$. We call this the *critical dataset size* for the transition.
- **Law 2.** Phase transitions in n that change the energy must *decrease* the energy and *increase* the learning coefficient.
- **Law 3.** Phase transitions in n that do not change the energy $E_\alpha = E_\beta$ must *decrease* the learning coefficient and *increase* the subleading term.

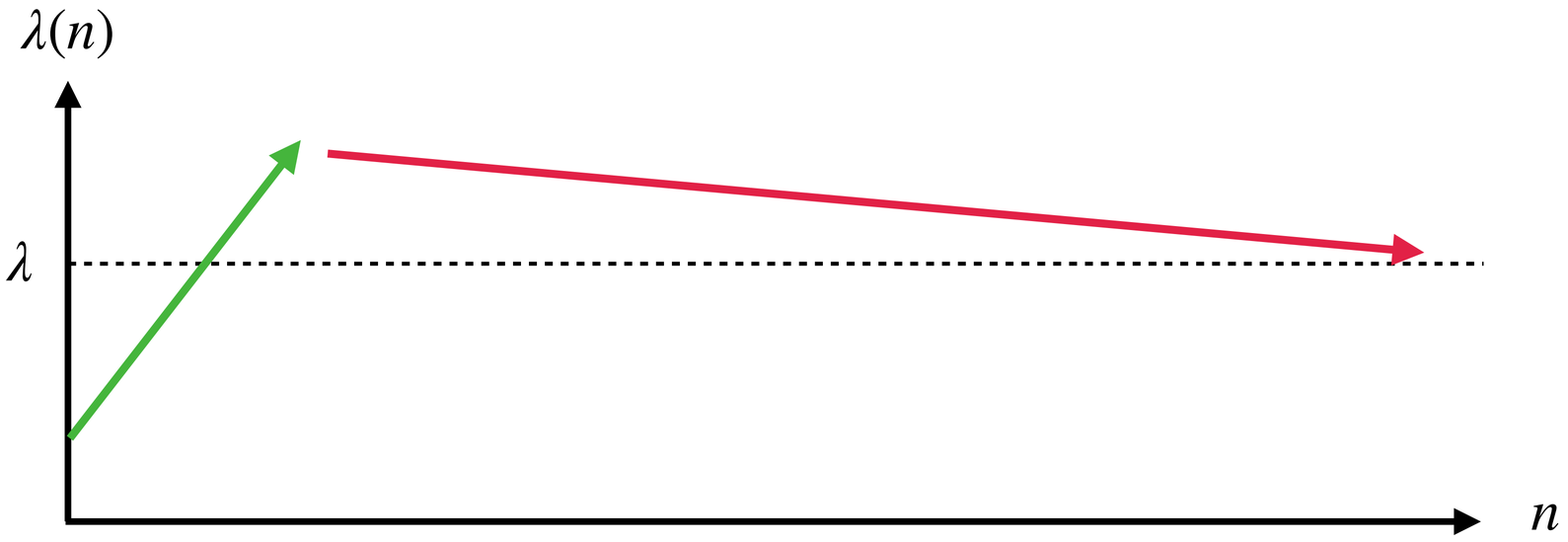
"Once the learning process reaches the set of optimal parameters, it undergoes transitions that *lower model complexity*"

Thermodynamics

$$F_n \approx \min_{\alpha} F_n(W_\alpha) \approx \min_{\alpha} \left[nE_\alpha + \lambda_\alpha \log n \right]$$

"The learning process produces *more accurate* models that are *more complex,* sacrificing extra bits in the model description for fewer errors"

"Once the learning process reaches the set of optimal parameters, it undergoes transitions that *lower model complexity*"



Behaviour of Generalisation Error

$$F_n \approx \min_{\alpha} F_n(W_\alpha) \approx \min_{\alpha} \left[nE_\alpha + \lambda_\alpha \log n \right]$$

- Suppose $E_\alpha > E_\beta$ and $\lambda_\alpha < \lambda_\beta$ and there is a phase transition $\alpha \longrightarrow \beta$.
- The generalisation error $\mathbb{E}[B_g]$ is affected by the phase transition, in one of two ways, depending on whether the generalisation errors of the two phases cross before or after the free energy:

$$E_\alpha + \frac{\lambda_\alpha}{n} = E_\beta + \frac{\lambda_\beta}{n} \quad \Longleftrightarrow \quad \frac{1}{n}\left(\lambda_\alpha - \lambda_\beta\right) = E_\beta - E_\alpha$$

$$\Longleftrightarrow \quad n = -\frac{\Delta\lambda}{\Delta E}$$

Thermodynamics

$$F_n \approx \min_\alpha F_n(W_\alpha) \approx \min_\alpha \left[nE_\alpha + \lambda_\alpha \log n\right]$$

- Suppose $E_\alpha > E_\beta$ and $\lambda_\alpha < \lambda_\beta$ and there is a phase transition $\alpha \longrightarrow \beta$.
- **Definition.** We say the transition $\alpha \longrightarrow \beta$ is *standard* if the generalisation error for phase β is lower than that of α at the transition. Otherwise it is a *lagging* transition.
- **Lemma.** If a transition is lagging then $n_{cr} < e$.
- Proof: With $r = -\dfrac{\Delta\lambda}{\Delta E}$ we have $r < n_{cr} = \mathcal{N}(r)$ implies $n_{cr} < \mathcal{N}(n_{cr})$ implies $n_{cr} \log n_{cr} < n_{cr}$ implies $n_{cr} < e$.
- Hence lagging transitions are essentially unobservable (or are they!).

Singular Learning Process

Gray Book, Section 7.6.

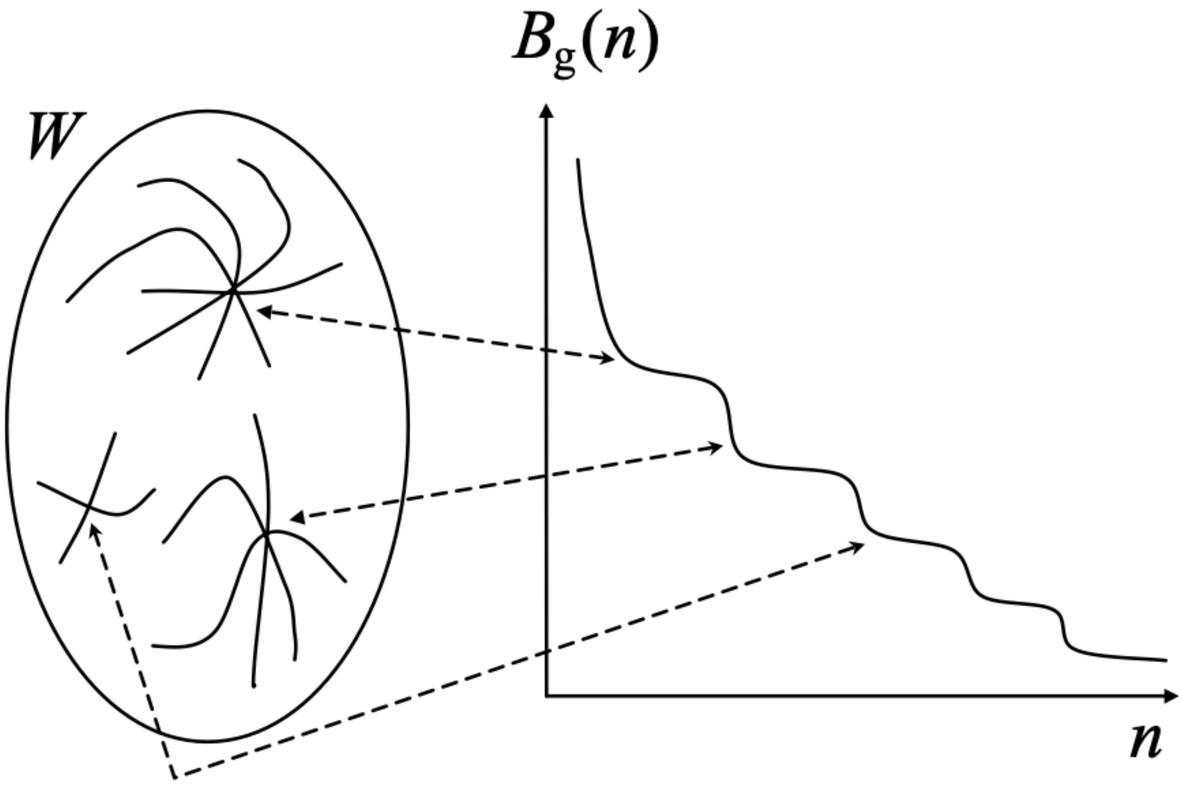
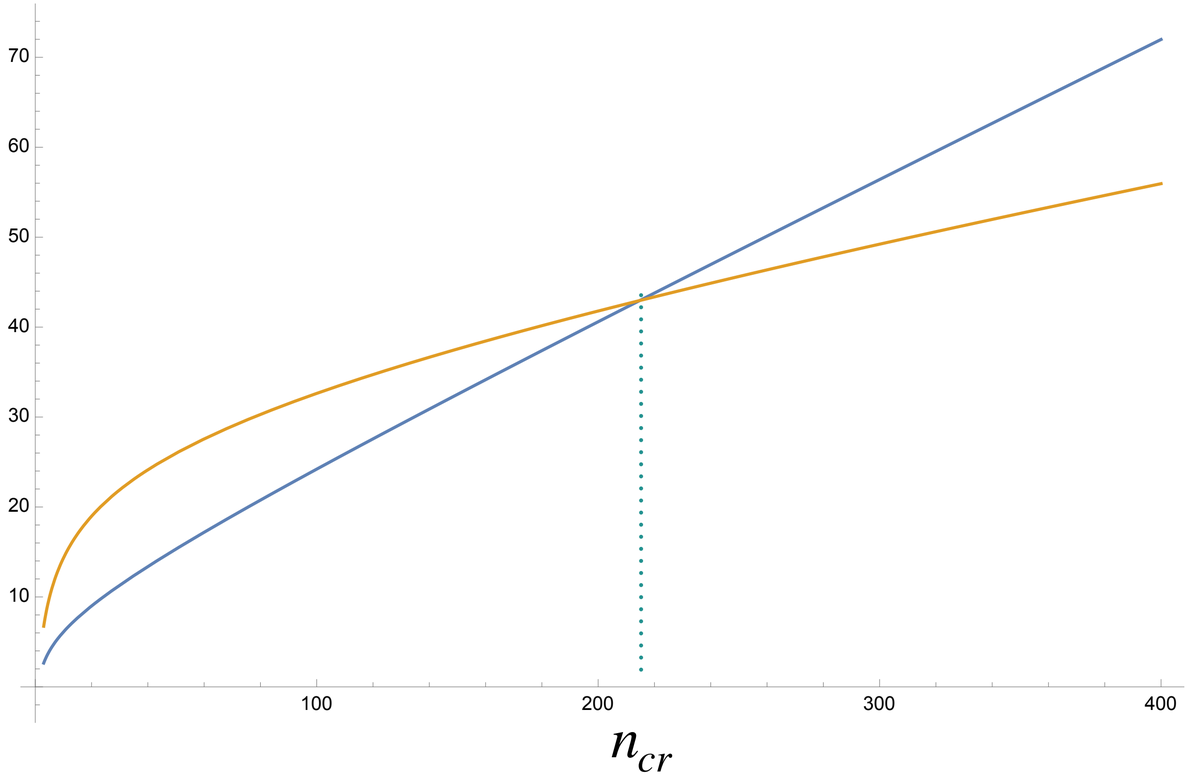


Fig. 7.6. Learning curve with singularities

Singular Learning Process



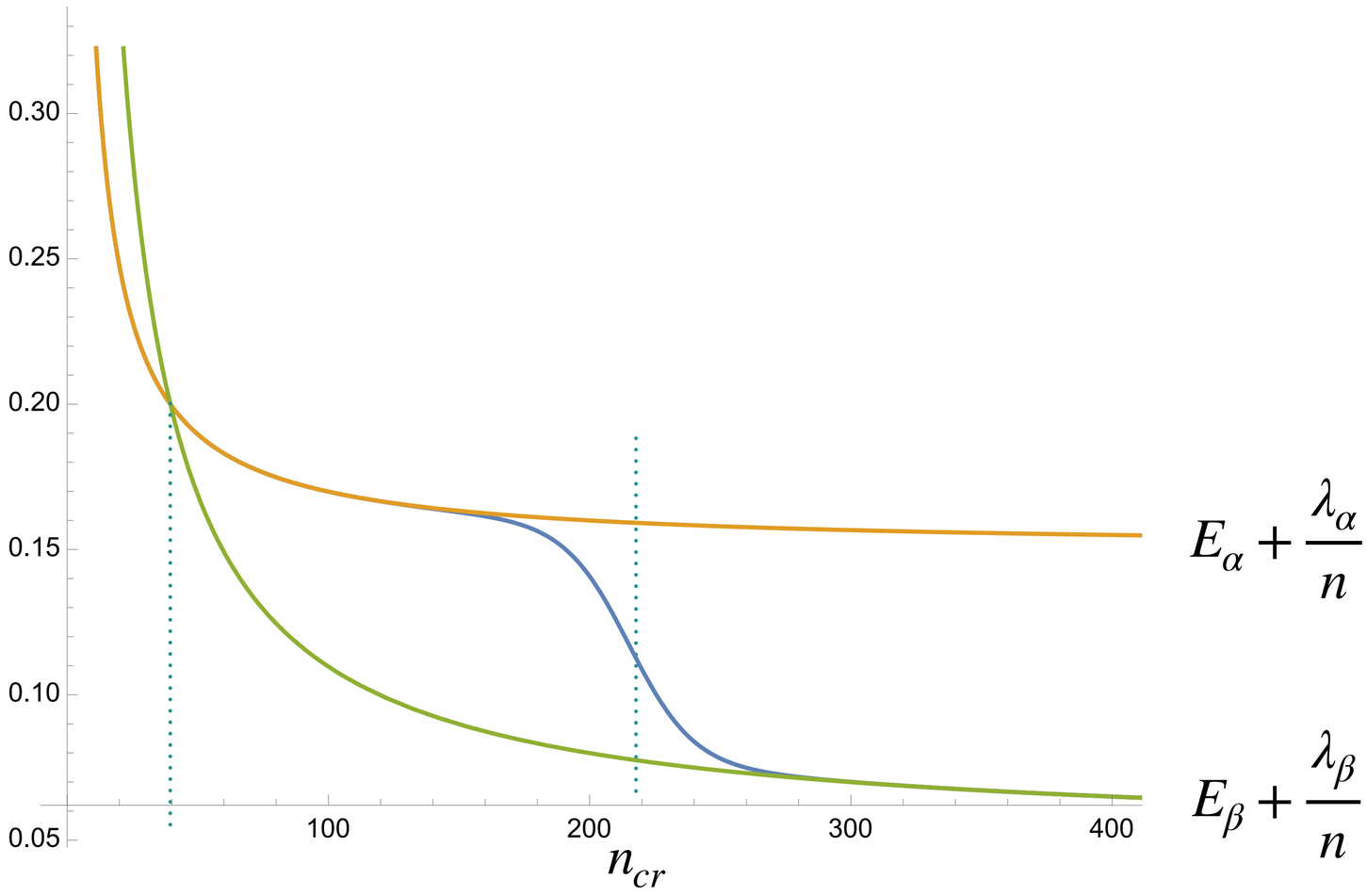
$$nE_\alpha + \lambda_\alpha \log n \qquad E_\alpha = 0.15, \quad \lambda_\alpha = 2$$

$$nE_\beta + \lambda_\beta \log n \qquad E_\beta = 0.05, \quad \lambda_\beta = 6$$

$$\Delta E = -0.1, \quad \Delta\lambda = 4$$

$$-\frac{\Delta\lambda}{\Delta E} = 40, \quad n_{cr} = 215$$

Singular Learning Process



$$E_\alpha = 0.15, \quad \lambda_\alpha = 2$$

$$E_\beta = 0.05, \quad \lambda_\beta = 6$$

$$\Delta E = -0.1, \quad \Delta\lambda = 4$$

$$-\frac{\Delta\lambda}{\Delta E} = 40, \quad n_{cr} = 215$$

Phase Structure

A Dummy's Theory of Scaling Laws

- We explore a simple model of *phase structure* meaning how the energy and complexity of the available phases vary with the phase index α. We assume there are a large number of phases, and take the indices α to be positive integers.
- Following Law 1 we assume that for $\alpha < \beta$ we have $E_\beta < E_\alpha$ and $\lambda_\beta > \lambda_\alpha$. For there to be an "unscreened" transition $\alpha \longrightarrow \alpha + 1$ we must have that the critical dataset size $n_{cr}(\alpha) = \mathcal{N}\left(-\frac{\Delta\lambda}{\Delta E}\right) = \mathcal{N}\left(-\frac{\lambda_{\alpha+1} - \lambda_\alpha}{E_{\alpha+1} - E_\alpha}\right)$ is an increasing function of α.
- One simple model is to take $E_\alpha = \frac{1}{\alpha}$ and $\lambda_\alpha = \alpha$. Then $\Delta\lambda = a$ and $\Delta E = \frac{-1}{\alpha(\alpha+1)}$ so $-\frac{\Delta\lambda}{\Delta E} = \alpha(\alpha+1) > 0$.

Phase Structure

A Dummy's Theory of Scaling Laws

- So our phase structure is $E_\alpha = \frac{1}{\alpha}$ and $\lambda_\alpha = \alpha$.
- To compute the dominant phase $\alpha(n)$ as a function of n we take $\alpha(n) = \text{argmin}_\alpha \left[nE_\alpha + \lambda_\alpha \log n \right]$, which we can compute by differentiating:

$$0 = \frac{d}{d\alpha}\left[nE_\alpha + \lambda_\alpha \log n \right] = \frac{-n}{\alpha^2} + \log n \Rightarrow \frac{n}{\log n} = \alpha^2$$

- Hence $\alpha(n) = \sqrt{\frac{n}{\log n}}$. This tells us the dominant phase at dataset size n, using which we can compute the overall behaviour of generalisation error.

Phase Structure

A Dummy's Theory of Scaling Laws

$$
\begin{aligned}
\mathbb{E}[B_g(n)] &= E_{\alpha(n)} + \frac{\lambda_{\alpha(n)}}{n} \\
&= \frac{1}{\alpha(n)} + \frac{\alpha(n)}{n} \\
&= \sqrt{\frac{\log n}{n}} + \frac{1}{n}\sqrt{\frac{n}{\log n}} \\
&= \sqrt{\frac{\log n}{n}}\left[1 + \frac{1}{\log n}\right] \\
&\approx \left(\frac{\log n}{n}\right)^{1/2}
\end{aligned}
$$