

Fast Rate Generalization Error Bounds: Variations on a Theme

Xuetong Wu¹ Jonathan H. Manton¹ Uwe Aickelin² Jingge Zhu¹

¹Department of Electrical and Electronic Engineering

²Department of Computing and Information Systems
University of Melbourne, Australia

IEEE Information Theory Workshop
Nov, 2022

Generalization in Machine Learning

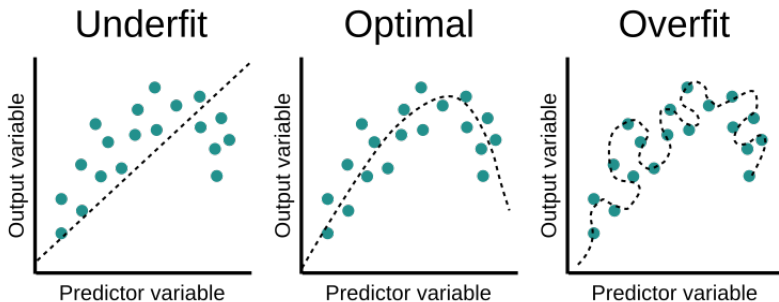


Figure: Overfitting and Underfitting¹

¹Source: www.educative.io/edpresso

Traditional Learning Theory

- Dataset $S := \{z_1, z_2, \dots, z_n\}$
- Instance: $z_i \in \mathcal{Z} : \mathcal{X} \times \mathcal{Y}$ (Supervised) or $\mathcal{X}$ (Unsupervised)
- Distributions: Z_i i.i.d. drawn from μ
- Output hypothesis: $w \in \mathcal{W}$
- Loss function $\ell : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}^+$
- Optimal hypothesis $w^* := \arg \min_{w \in \mathcal{W}} \mathbb{E}_{Z \sim \mu} [\ell(w^*, Z)]$
- ERM hypothesis $w_{\text{ERM}} := \arg \min_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n \ell(w, z_i)$
- Generalization Error:

$$\mathcal{E}(w, S) := \mathbb{E}_{Z \sim \mu} [\ell(w, Z)] - \frac{1}{n} \sum_{i=1}^n \ell(w, z_i)$$

- Unexpected excess risk: $r(w, z) = \ell(w, z) - \ell(w^*, z)$
- Excess risk:

$$R(w) := \mathbb{E}_{Z \sim \mu} [\ell(w, Z)] - \mathbb{E}_{Z \sim \mu} [\ell(w^*, Z)]$$

- Xu & Raginsky [NIPS'17]: $\mathbb{E}[\mathcal{E}(W, S)] \leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}$
 - **Issue:** $I(W; S)$ can be arbitrarily large for deterministic algorithms such as ERM.
 - Negrea et al. [NIPS'19]: $\mathbb{E}[\mathcal{E}(W, S)] \leq \mathbb{E}_{u \sim U} \sqrt{\frac{2\sigma^2}{m} I(W; \mathcal{S}_u)}$
 - Bu et al. [ISIT'19]: $\mathbb{E}[\mathcal{E}(W, S)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(W; Z_i)}$
 - Thomas & Lydia [COLT'20]: $\mathbb{E}[\mathcal{E}(W, S)] \leq \sqrt{\frac{2\sigma^2}{n} I(W; J | \mathcal{S}^{2n})}$
 - **Another Issue:** sometimes the information-theoretic bounds do not capture the true convergence rate even in a very simple Gaussian mean estimation problem.

Example: Gaussian Mean Estimation

- Data: $Z_i \sim \mathcal{N}(\mu, \sigma_N^2)$
- Loss: $\ell(w, z_i) = (w - z_i)^2$
- ERM: $W_{\text{ERM}} = \frac{1}{n} \sum_{i=1}^n Z_i \sim \mathcal{N}(\mu, \frac{\sigma_N^2}{n})$
- Loss distribution: $\ell(W, Z_i) \sim \frac{n+1}{n} \sigma_N^2 \chi_1^2 \rightarrow \sqrt{2\sigma_W^4}$ -sub-Gaussian under $P_W \otimes \mu$
where $\sigma_W^2 = \frac{n+1}{n} \sigma_N^2$ for any i
- Mutual information: $I(W; Z_i) = \frac{1}{2} \log \frac{n}{n-1}$
- Bounds:

$$\mathbb{E}_{WS} [\mathcal{E}(W, S)] \leq \frac{\sigma_N^2}{n} \sum_{i=1}^n \sqrt{2 \frac{(n+1)^2}{n^2} \log \frac{n}{n-1}} \sim \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

- True Generalization Error:

$$\mathbb{E}_{WS} [\mathcal{E}(W_{\text{ERM}}, S)] = \frac{2\sigma_N^2}{n} \sim \mathcal{O}\left(\frac{1}{n}\right)$$

Main Results: Change on the Assumption

Theorem (Bu et al. (ISIT'19))

Suppose that $\ell(W, Z)$ is σ -sub-Gaussian under the distribution $P_W \otimes \mu$, then:

$$\mathbb{E}_{W\mathcal{S}_n} [\mathcal{E}(W, \mathcal{S}_n)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(W; Z_i)}$$

Theorem

Suppose that $r(W, Z)$ is σ -sub-Gaussian under the distribution $P_W \otimes \mu$, then:

$$\mathbb{E}_{W\mathcal{S}_n} [\mathcal{E}(W, \mathcal{S}_n)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(W; Z_i)}$$

Recall: $r(w, z) = \ell(w, z) - \ell(w^*, z)$.

Theorem

Suppose that $r(W, Z)$ is σ -sub-Gaussian under the distribution $P_W \otimes \mu$, then:

$$\mathbb{E}_{W\mathcal{S}_n} [\mathcal{E}(W, \mathcal{S}_n)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(W; Z_i)}$$

To achieve fast rate:

- The output hypothesis of the learning algorithm must be “good” enough compared to the optimal hypothesis w^* ;
- We encode the notion of goodness in terms of the CGF by controlling $r(w, z)$.

Example Continued: Gaussian Mean Estimation

- $Z_i \sim \mathcal{N}(\mu, \sigma_N^2)$
- $r(w, z_i) = (w - z_i)^2 - (\mu - z_i)^2$
- Expected Excess risk: $\mathbb{E}_{P_W \otimes \mu}[r(w, z)] = \frac{\sigma_N^2}{n}$
- CGF: $\log \mathbb{E}_{P_W \otimes \mu} [e^{\eta(r(W, Z) - \mathbb{E}[r(W, Z)])}] \leq \frac{2\eta^2 \sigma_N^4}{n}$ for any η : $r(W, Z)$ is $\sqrt{\frac{4\sigma_N^4}{n}}$ -sub-Gaussian.
- The MI bound:

$$\mathbb{E}_{WS} [\mathcal{E}(W, S)] \leq \frac{\sigma_N^2}{n} \sum_{i=1}^n \sqrt{\frac{4}{n} \log \frac{n}{n-1}} \approx \frac{2\sigma_N^2}{n} \sim o\left(\frac{1}{n}\right)$$

- True Generalization Error:

$$\mathbb{E}_{WS} [\mathcal{E}(W_{\text{ERM}}, S)] = \frac{2\sigma_N^2}{n} \sim o\left(\frac{1}{n}\right)$$

Once we have made the assumption on the excess risk, we observe that:

- Subgaussian on ℓ : $\sqrt{2\sigma_W^4}$ -sub-Gaussian
- Subgaussian on r : $\sqrt{\frac{4\sigma_N^4}{n}}$ -sub-Gaussian

The CGF under r can converge to zero with n increasing but not the case with ℓ
→ difference in rate.

Theorem (Fast Rate with Sub-Gaussian Condition)

Let $\hat{R}(w, S) := \frac{1}{n} \sum_{i=1}^n r(w, z_i)$. Assume that $r(W, Z)$ is σ -sub-Gaussian under the distribution $P_W \otimes \mu$:

$$\mathbb{E}_{WS}[\mathcal{E}(W, S)] \leq \frac{1 - a_\eta}{a_\eta} \mathbb{E}_{WS}[\hat{R}(W, S)] + \frac{1}{n\eta a_\eta} \sum_{i=1}^n I(W; Z_i).$$

for any $0 < \eta < \frac{2\mathbb{E}_{P_W \otimes \mu}[r(W, Z_i)]}{\sigma^2}$ and $a_\eta = 1 - \frac{\eta\sigma^2}{2\mathbb{E}_{P_W \otimes \mu}[r(W, Z_i)]}$.

However, both η and a_η depend on $\mathbb{E}_{P_W \otimes \mu}[r(W, Z)]$ and σ^2 , which potentially depend on n as well.

New conditions needed!

Expected (η, c) -Central Condition

Definition (Expected (η, c) -Central Condition)

Let $\eta > 0$ and $0 < c \leq 1$ be two constants. We say that $(\mu, \ell, \mathcal{W}, \mathcal{A})$ satisfies the expected (η, c) -central condition if the following inequality holds for the optimal hypothesis w^* :

$$\log \mathbb{E}_{P_W \otimes \mu} \left[e^{-\eta r(W, Z)} \right] \leq -c\eta \mathbb{E}_{P_W \otimes \mu} [r(W, Z)].$$

Conventional η -central condition [Van Erven et al., 2015, Def. 3.1]:

$$\log \mathbb{E}_{P_W \otimes \mu} \left[e^{-\eta r(W, Z)} \right] \leq 0, \eta > 0$$

- The RHS of (η, c) -CC is negative and has a tighter control than η -CC of the tail behaviour.
- Roughly speaking, this condition ensures that for all $\eta' \leq \eta$, the probability that w outperforms w^* by more than L is exponentially small in $L + c\mathbb{E}[R(W)]$.

Closely related works on fast rate

- Bernstein's Condition [Van Erven et al., 2015]: “easiness” of the learning problem with second moment.
- Central condition and witness condition [Grünwald et al., 2020]: “easiness” of the learning problem with MGF and observability.

Corollary

Let $\beta \in [0, 1]$ and $B \geq 1$. For a learning tuple $(\mu, \ell, \mathcal{W}, \mathcal{A})$, we say that the **Bernstein condition** holds if the following inequality holds for the optimal hypothesis w^* :

$$\mathbb{E}_{P_W \otimes \mu} [r(W, Z)^2] \leq B (\mathbb{E}_{P_W \otimes \mu} [r(W, Z)])^\beta.$$

Then, if $\beta = 1$ and $r(w, z_i)$ is bounded by $-b$ with some $b > 0$ for all w and z_i , the learning tuple also satisfies $(\min(\frac{1}{b}, \frac{1}{2B(e-2)}), \frac{1}{2})$ -central condition.

η -Central with Witness Condition

- η -central condition:

$$\mathbb{E}_{P_W \otimes \mu} \left[e^{-\eta(r(W, Z))} \right] \leq 1.$$

- (u, c) -witness condition for $u > 0$ and $c \in (0, 1]$:

$$\mathbb{E}_{P_W \otimes \mu} [r(W, Z) \cdot \mathbf{1}_{\{r(W, Z) \leq u\}}] \geq c \mathbb{E}_{P_W \otimes \mu} [r(W, Z)],$$

Corollary

If the learning tuple satisfies both η -central condition and (u, c) -witness condition, then the learning tuple also satisfies the $(\eta', \frac{c - \frac{c\eta'}{\eta}}{\eta' u + 1})$ -central condition for any $0 < \eta' < \eta$.

(η, c) -Central Condition Theorem

Theorem (Fast Rate with (η, c) -central condition)

Assume the learning tuple $(\mu, \ell, \mathcal{W}, \mathcal{A})$ satisfies the (η, c) -central condition for some constants $\eta > 0$ and $0 < c \leq 1$. Then, for all $\eta' \in (0, \eta]$, it holds that,

$$\mathbb{E}_{WS}[\mathcal{E}(W, S)] \leq \frac{1-c}{c} \mathbb{E}_{WS}[\hat{\mathcal{R}}(W, S)] + \frac{1}{c\eta'n} \sum_{i=1}^n I(W; Z_i).$$

For algorithms such as ERM, the first term will be negative and hence,

$$\mathbb{E}_{WS}[\mathcal{E}(W, S)] \leq \frac{1}{c\eta'n} \sum_{i=1}^n I(W; Z_i)$$

achieving $O(\frac{1}{n})$ for $I(W; Z_i) \sim O(\frac{1}{n})$.

Example Continued: Gaussian Mean Estimation

For any n ,

$$\log \mathbb{E}_{P_W \otimes \mu} \left[e^{-\eta r(W, Z)} \right] \leq \frac{2\eta^2 \sigma_N^4 - \eta \sigma_N^2}{n} \leq -c\eta \frac{\sigma_N^2}{n}.$$

This learning problem satisfy the (η, c) -central condition for:

- $0 < \eta < \frac{1}{2\sigma_N^2}$
- $c \leq 1 - 2\eta\sigma_N^2$

Taking $\eta = \frac{1}{4\sigma_N^2}$ and $c = \frac{1}{2}$, the bound becomes

$$\frac{1-c}{c} \mathbb{E}_{WS}[\hat{R}(W, S)] + \frac{1}{c\eta'n} \sum_{i=1}^n I(W; Z_i) = \frac{3\sigma_N^2}{n},$$

arriving at the fast rate $O(1/n)$.

Extension: Excess risk with Regularized ERM

- Algorithm [Koren & Y. Levy, 2015]:

$$w_{\text{RERM}} = \arg \min_{w \in \mathcal{W}} \hat{L}(w, S) + \frac{\lambda}{n} g(w),$$

where $g : \mathcal{W} \rightarrow \mathbb{R}$ denotes the regularizer.

- Assume Boundness: $|g(w_1) - g(w_2)| \leq B$ for any w_1 and w_2 in $\mathcal{W}$
- Bounds for excess risk:

$$\mathbb{E}_{\mathcal{W}}[R(w_{\text{RERM}})] \leq \frac{\lambda B}{cn} + \frac{1}{c\eta' n} \sum_{i=1}^n I(w_{\text{RERM}}; Z_i).$$

Extension: Intermediate Rate

Definition ((ν, c)-Central Condition)

We say that $(\mu, \ell, \mathcal{W}, \mathcal{A})$ satisfies the (η, c) -central condition up to some $\epsilon > 0$ if the following inequality holds for the optimal hypothesis w^* :

$$\log \mathbb{E}_{P_{W \otimes \mu}} \left[e^{-\eta r(W, Z)} \right] \leq -c\eta \mathbb{E}_{P_{W \otimes \mu}} [r(W, Z)] + \eta \epsilon.$$

Let $\nu : [0, \infty) \rightarrow [0, \infty)$ is a bounded and non-decreasing function satisfying $\nu(\epsilon) > 0$ for all $\epsilon > 0$. We say that $(\mu, \ell, \mathcal{W}, \mathcal{A})$ satisfies the (ν, c) -central condition if for all $\epsilon \geq 0$, (η, c) -central condition up to ϵ is satisfied with $\eta = \nu(\epsilon)$.

Extension: Intermediate Rate

- Assume (ν, c) -Central Condition holds,

$$\mathbb{E}_{WS}[\mathcal{E}(W, S)] \leq \frac{1-c}{c} \mathbb{E}_{WS}[\hat{R}(W, S)] + \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{\nu(\epsilon)c} I(W; Z_i) + \frac{\epsilon}{c} \right).$$

- In particular, if $\nu(\epsilon) \asymp \epsilon^{1-\beta}$ for some $\beta \in [0, 1]$:

$$\mathbb{E}_{WS}[\mathcal{E}(W, S)] \leq \frac{1-c}{c} \mathbb{E}_{WS}[\hat{R}(W, S)] + \frac{2}{nc} \sum_{i=1}^n I(W; Z_i)^{\frac{1}{2-\beta}}.$$

corresponding to the results under Bernstein's condition [Hanneke 2016, Mhammedi 2019, Grünwald 2021].

Table: Technical Conditions Comparisons

Condition	Key Inequality
(η, c) -Central	$\log \mathbb{E} [e^{-\eta r(W, Z)}] \leq -c\eta \mathbb{E}[r(W, Z)]$
Bernstein with $\beta = 1$	$\log \mathbb{E} [e^{-\eta r(W, Z)}] \leq -\frac{1}{2}\eta \mathbb{E}[r(W, Z)]$
Central + Witness	$\log \mathbb{E} [e^{-\eta r(W, Z)}] \leq -\frac{1}{c_\mu}\eta \mathbb{E}[r(W, Z)]$
Central Only	$\log \mathbb{E} [e^{-\eta r(W, Z)}] \leq 0$
Sub-Gaussian	$\log \mathbb{E} [e^{-\eta r(W, Z)}] \leq -\eta \mathbb{E}[r(W, Z)] + \frac{\eta^2 \sigma^2}{2}$

- Non-trivial verification for (η, c) -central condition;
- More learning tasks and algorithms that satisfy (η, c) -condition;
- Fast rate other than ERM, RERM.

References I



Bu, Y., Zou, S., and Veeravalli, V. V. (2020).
Tightening mutual information-based bounds on generalization error.
IEEE Journal on Selected Areas in Information Theory, 1(1):121–130.



Grünwald, P., Steinke, T., and Zakyntinou, L. (2021).
Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds
that handle general vc classes.
arXiv preprint arXiv:2106.09683.



Grünwald, P. D. and Mehta, N. A. (2020).
Fast rates for general unbounded loss functions: From erm to generalized bayes.
J. Mach. Learn. Res., 21:56–1.

References II



Hanneke, S. (2016).

Refined error bounds for several learning algorithms.

The Journal of Machine Learning Research, 17(1):4667–4721.



Koren, T. and Levy, K. (2015).

Fast rates for exp-concave empirical risk minimization.

Advances in Neural Information Processing Systems, 28.



Mhammedi, Z., Grunwald, P. D., and Guedj, B. (2019).

Pac-bayes un-expected bernstein inequality.

arXiv preprint arXiv:1905.13367.

References III



Negrea, J., Haghifam, M., Khisti, A., Dziugaite, G. K., and Roy, D. M. (2019).

Information-theoretic generalization bounds for sgld via data-dependent estimates.

arXiv, (NeurIPS).



Steinke, T. and Zakynthinou, L. (2020).

Reasoning about generalization via conditional mutual information.

In *Conference on Learning Theory*, pages 3437–3452. PMLR.



Van Erven, T., Grunwald, P., Mehta, N. A., Reid, M., Williamson, R., et al. (2015).

Fast rates in statistical and online learning.



Xu, A. and Raginsky, M. (2017).

Information-theoretic analysis of generalization capability of learning algorithms.

In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2521–2530.