

A Compressibility Approach to Generalization and (No) Memorization

Abdellatif Zaidi[†]

[†]Université Gustave-Eiffel and Paris Research Center, Huawei Technologies France

Inria, Université Côte d'Azur, February 2026

1 Part I: The Problem of Generalization

- Fixed-size compressibility
- Variable-size compressibility
- Unification and new bounds

2 Part II: The Problem of Memorization

- Is Memorization Necessary ?
- Implications

Preliminaries - Generalization Error

Data: For an unknown distribution μ on $\mathcal{Z}$,

- Dataset: n i.i.d. samples $S = \{Z_i\}_{i=1}^n \sim \mu^{\otimes n}$
- Example: $Z_i = (X_i, Y_i) \sim \mu$

Learning algorithm: mapping $\mathcal{A}$ from $S \in \mathcal{Z}^n$ to a hypothesis $W \in \mathcal{W}$.

Loss function: $\ell : \mathcal{Z} \times \mathcal{W} \rightarrow \mathbb{R}_+$.

E.g., for binary classification:

$z = (x, y)$, $x \in \mathbb{R}^d$, $y \in \{-1, 1\}$, $\ell(z, w) := \mathbb{1}_{yf(x, w) < 0}$ (0-1 loss) with f decision function.

Population risk:

$$\mathcal{L}(w) := \mathbb{E}_{Z \sim \mu}[\ell(Z, w)]$$

Empirical risk:

$$\hat{\mathcal{L}}(S, w) := \frac{1}{n} \sum_{i=1}^n \ell(z_i, w)$$

Generalization error:

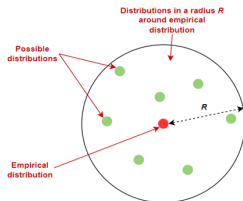
$$\text{gen}(S, w) := \mathcal{L}(w) - \hat{\mathcal{L}}(S, w)$$

Generalization error depends on:

- Loss function
- Data distribution
- Size of training dataset
- Learning algorithm

Related to algorithmic stability & robustness:

- Uniform stability
- Average stability



Exact analysis out of reach - resort to bounds:

- Tail bounds
- In-expectation bounds

Compressibility framework

- **Idea:** - Find a proxy algorithm $\hat{\mathcal{A}}(S, W) = \hat{W} \in \hat{\mathcal{W}} \subseteq \mathcal{W}$ such that: (i) $\hat{\mathcal{A}}$ performs close to the original $\mathcal{A}$ generalization-error wise and (ii) whose complexity (e.g., $|\hat{\mathcal{W}}|$) is minimal.
 - That (minimal) complexity of the proxy $\hat{\mathcal{A}}$ is what controls the generalization error of the original $\mathcal{A}$!

[SGRS, COLT 2022] Milad Sefidgaran, Amin Gohari, Gaël Richard, Umut Şimşekli, "Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms", 35th Conference on Learning Theory (COLT) 2022

Compressibility framework

- **Idea:** - Find a proxy algorithm $\hat{\mathcal{A}}(S, W) = \hat{W} \in \hat{\mathcal{W}} \subseteq \mathcal{W}$ such that: (i) $\hat{\mathcal{A}}$ performs close to the original $\mathcal{A}$ generalization-error wise and (ii) whose complexity (e.g., $|\hat{\mathcal{W}}|$) is minimal.
 - That (minimal) complexity of the proxy $\hat{\mathcal{A}}$ is what controls the generalization error of the original $\mathcal{A}$!
- **Rate-Distortion Theoretic Approach:** Random **block** coding. Consider m independent training datasets $S^m = (S_1, \dots, S_m)$ and associated hypotheses $W^m = (W_1, \dots, W_m)$, with $W_i = \mathcal{A}(S_i)$
 - Fix a **compression hypothesis book** $\mathcal{H}_m = \{\hat{\mathbf{w}}_j : j \in [e^{mR}]\} \subseteq \hat{\mathcal{W}}^m$, with $\hat{\mathbf{w}}_j = (\hat{w}_{j,1}, \dots, \hat{w}_{j,m}) \in \hat{\mathcal{W}}^m$ and $\hat{\mathcal{A}}_m : \mathcal{S}^m \times \mathcal{W}^m \rightarrow \hat{\mathcal{W}}^m$
 - Choose a distortion measure $\tilde{d}_m : \mathcal{S}^m \times \mathcal{W}^m \times \hat{\mathcal{W}}^m \rightarrow \mathbb{R}$

$$\tilde{d}_m(w^m, \hat{w}^m; s^m) := \frac{1}{m} \sum_{i=1}^m (\text{gen}(s_i, w_i) - \text{gen}(s_i, \hat{w}_i))$$

- w^m is **covered** by $\mathcal{H}_m$ if there exists at least one $j \in [e^{mR}]$ such that $\tilde{d}_m(w^m, \hat{w}^m; s^m) < \epsilon$. If so, $\hat{\mathcal{A}}_m^*(s^m, w^m; \mathcal{H}_m) = \hat{\mathbf{w}}_j$, with $j = \arg \min_{j \in [l_m]} \tilde{d}_m(w^m, \hat{\mathbf{w}}_j; s^m)$

[SGRS, COLT 2022] Milad Sefidgaran, Amin Gohari, Gaël Richard, Umut Şimşekli, "Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms", 35th Conference on Learning Theory (COLT) 2022

Fixed-size compressibility framework

- **Error event:** when $(S^m, \mathcal{A}(S^m))$ is not '**covered**':

$$\mathcal{E}_m(\mathcal{H}_m, \epsilon) := \left\{ \min_{j \in [l_m]} \tilde{d}_m(\mathcal{A}(S^m), \hat{\mathbf{w}}_j; S^m) > \epsilon \right\}.$$

- **In-expectation bound:** Minimum R such that

$$\lim_{m \rightarrow \infty} \mathbb{P}_{(S, W) \otimes m}(\mathcal{E}_m(\mathcal{H}_m, \epsilon)) = 0$$

- **Tail bound:** Minimum R such that

$$\lim_{m \rightarrow \infty} \left[-\frac{1}{m} \log \left(\mathbb{P}_{(S, W) \otimes m}(\mathcal{E}_m(\mathcal{H}_m, \epsilon)) \right) \right] \geq \log(1/\delta)$$

Hence, $\mathbb{P}_{(S, W) \otimes m}(\mathcal{E}_m(\mathcal{H}_m, \epsilon)) \lesssim \delta^m$.

Bounds on the expectation of generalization error

Bounds on the expectation of generalization error

Fix $\epsilon \in \mathbb{R}$. If $\ell(Z, w)$ is σ -subgaussian $\forall w \in \mathcal{W}$, then

$$\mathbb{E}[\text{gen}(S, W)] \leq \sqrt{\frac{2\sigma^2 \mathfrak{RD}(\epsilon; P_{S, W})}{n}} + \epsilon.$$

where

$$\mathfrak{RD}(\epsilon; P_{S, W}) = \inf_{P_{\hat{W}|S}} I(S; \hat{W}), \quad \text{such that} \quad \mathbb{E}[\text{gen}(S, W) - \text{gen}(S, \hat{W})] \leq \epsilon.$$

The approach can be easily extended to bound $|\mathbb{E}[\text{gen}(S, W)]|$ with a similar term $\Rightarrow$ recovers e.g. Xu & Raginsky (2018) by letting $\epsilon = 0$ and $P_{\hat{W}|S} = P_{W|S}$:

- $I(S; W)$ is an upper bound on the *lossless* compressibility of the learning algorithm

Tail bound on the generalization error

Tail bounds on the generalization error

If $\forall w \in \mathcal{W}$, $\ell(Z, w)$ is σ -subgaussian, then for every algorithm $\mathcal{A}(S)$ and every $\epsilon \in \mathbb{R}$, then with probability $1 - \delta$

$$\text{gen}(S, W) \leq \sqrt{\frac{2\sigma^2(R_p(\delta, \epsilon) + \log(1/\delta))}{n}} + \epsilon.$$

where

$$R_p(\delta, \epsilon) := \sup_{Q: D_{KL}(Q \| P_{S, W}) \leq \log(1/\delta)} \mathfrak{RD}(\epsilon; Q)$$

$$\text{and } \mathfrak{RD}(\epsilon; Q) = \inf_{P_{\hat{W}|S}} I(S; \hat{W}) \text{ such that } \mathbb{E}[\text{gen}(S, W) - \text{gen}(S, \hat{W})] \leq \epsilon.$$

★ The obtained bounds are **data-independent**: they depend on the data via $P_{S, W}$, not the given training sample S at hand!

Variable-size compressibility framework

- Define the distortion measures $d: \mathcal{S} \times \mathcal{W} \times \hat{\mathcal{W}} \rightarrow \mathbb{R}$ and $d_m: \mathcal{S}^m \times \mathcal{W}^m \times \hat{\mathcal{W}}^m \rightarrow \mathbb{R}$ as

$$d(w, \hat{w}; s) := \text{gen}(s, w)^2 - \text{gen}(s, \hat{w})^2, \quad d_m(w^m, \hat{w}^m; s^m) := \frac{1}{m} \sum_{i \in [m]} d(w_i, \hat{w}_i; s_i).$$

Variable-size compressibility definition

$\mathcal{A}(S)$ is $(R_{S,W}, \epsilon, \delta)$ -compressible for some $\{R_{s,w}\}_{(s,w) \in \mathcal{S} \times \mathcal{W}}$, where $R_{s,w} \in \mathbb{R}^+$ and some $\epsilon \in \mathbb{R}$ and $\delta \in \mathbb{R}^+$, if $\exists \{\mathcal{H}_m\}_{m \in \mathbb{N}}$, $\mathcal{H}_m := \{\hat{\mathbf{w}}[j], j \in \mathbb{N}^*\}$, such that

$$\lim_{m \rightarrow \infty} \left[-\frac{1}{m} \log \mathbb{P}_{(S,W) \otimes m} \left(\min_{j \leq e^{\sum_{i \in [m]} R_{S_i, W_i}}} d_m(W^m, \hat{\mathbf{w}}[j]; S^m) > \epsilon \right) \right] \geq \log(1/\delta).$$

- The number of indices in $\mathcal{H}_m$ searched for a suitable *covering* of (s^m, w^m) depends on (s^m, w^m) .

Tail bounds on the generalization error

Variable-size compressibility of $\mathcal{A}(S)$

If for every $\nu_{S,W}$ s.t. $D_{KL}(\nu_{S,W} \| P_{S,W}) \leq \log(1/\delta)$,

$$\inf_{p_{\hat{W}|S} \in \mathcal{Q}(\nu_{S,W})} \inf_{q_{\hat{W}}} \left\{ D_{KL}(p_{\hat{W}|S} \nu_S \| q_{\hat{W}} \nu_S) - D_{KL}(\nu_{S,W} \| P_{W|S} \nu_S) \right\} \leq \mathbb{E}_{\nu_{S,W}} [R_{S,W}].$$

where $\mathcal{Q}(\nu_{S,W})$ stands for the set of distributions $p_{\hat{W}|S}$ that satisfy $\mathbb{E}_{\nu_{S,W} p_{\hat{W}|S}} [\text{gen}(S, W)^2 - \text{gen}(S, \hat{W})^2] \leq \epsilon$; and $\ell(Z, w)$ is σ -subgaussian $\forall w \in \mathcal{W}$, then with probability $1 - \delta$

$$\text{gen}(S, W) \leq \sqrt{\frac{4\sigma^2 \left(R_{S,W} + \log(\sqrt{2n}/\delta) \right)}{2n - 1}} + \epsilon.$$

Example: Set $\epsilon = 0$. Choice $R_{S,W} := \log \frac{dP_{W|S}}{dP}(W)$ recovers disintegrated PAC-Bayes bound of Catoni (2007)

$$\text{gen}(S, W) \leq \sqrt{4\sigma^2 \left(\log \frac{dP_{W|S}}{dP}(W) + \log(\sqrt{2n}/\delta) \right)} / (2n - 1). \quad (\text{Can also recover McAllester (1998)})$$

Novel bounds

- Let $T := [t_1 : t_2] \subseteq \mathbb{N}$ $\text{gen}(S, W_S^T) := \frac{1}{t_2 - t_1} \sum_{t \in T} \text{gen}(S, W_S^t)$.
- Assume that the loss is $\mathfrak{L}$ -Lipschitz, i.e. $\forall z \in \mathcal{Z}, w, \hat{w} \in \mathcal{W}: |\ell(z, w) - \ell(z, \hat{w})| \leq \mathfrak{L}\rho(w, \hat{w})$. Define the rate-distortion of the optimization trajectory T for any $S = s$ and $P_{W^T|s}$ as

$$\mathfrak{RD}(s, \epsilon; P_{W^T|s}) := \inf_{P_{\hat{W}^T|W^T, s}} I(W^T; \hat{W}^T), \quad \text{s.t.} \quad \mathbb{E}[\rho(W^T, \hat{W}^T)] \leq \epsilon.$$

where $\rho(W^T, \hat{W}^T) := \frac{1}{t_2 - t_1} \sum_{t \in T} \rho(w^t, \hat{w}^t)$ is the average distortion along the iterations.

General tail bound in terms of RD of the trajectory

For $\mathfrak{L}$ -Lipschitz loss that takes value in $[0, 1]$ and for any $\epsilon \in \mathbb{R}$, with probability $1 - \delta$,

$$\mathbb{E}_{W^T \sim P_{W^T|S}} [\text{gen}(S, W^T)] \leq \sqrt{\frac{\mathfrak{RD}(S, \epsilon; P_{W^T|S}) + \log(\sqrt{2n}M/\delta)}{2n - 1}} + 4\mathfrak{L}\epsilon.$$

- M is a coupling constant, defined in Şimşekli et al (2020), upper bounded by $I_\infty(S; W^T)$.

Wrap-up on Part I

- A unifying variable-size compressibility framework is introduced.
- The framework recovers, and gives a new interpretation of, most of existing seemingly unrelated approaches
 - Information-theoretic bounds (e.g. Xu & Raginsky (2017))
 - PAC-Bayes bounds (e.g. Rivasplata et al. (2020))
 - Compression bounds (e.g. Arora et al. (2018))
 - Intrinsic-dimension bounds (e.g. Şimşekli et al. (2020))
- The framework allows to establish **novel** bounds including
 - A new lossy PAC-Bayes bound,
 - A new bound in terms of the rate-distortion of the process.
- Similar tail bounds on the conditionally expected generalization error (e.g. with respect to $P_{W|S}$) and in-expectation bounds can be established; or to include Rényi divergence terms instead of KL-divergence.

[SZ, TIT 2024] M. Sefidgaran and A. Zaidi, Data-dependent generalization bounds via variable size compressibility, TIT 2024, Vol. 70, no. 9, pp. 6572-6595, Sept. 2024

The Problem of Memorization

How much a “good” algorithm must reveal about its training data ?

The Problem of Memorization

How much a “good” algorithm must reveal about its training data ?

- **Complex relationship between generalization and memorization !**

- Benefits of limited memorization for strong generalization

- D. Russo and J. Zou. How much does your data exploration overfit? Controlling bias via information usage. 2015. arXiv: 1511.05219.
 - R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. “Learners that Use Little Information”. In: Algorithmic Learning Theory. 2018, pp. 25–55.

- But, modern overparametrized DNNs often achieve high accuracy despite memorizing a significant part of the training data !

- N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, “Quantifying memorization across neural language models”’. arXiv preprint arXiv:2202.07646 (2022).
 - V. Feldman and C. Zhang. “What neural networks memorize and why: Discovering the long tail via influence estimation”. Advances in Neural Information Processing Systems 33 (2020), pp. 2881–2891.

- Memorization might even be *necessary* for generalization !

- V. Feldman. “Does learning require memorization ? a short tale about a long tail”’. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. 2020, pp. 954–959.
 - G. Brown, M. Bun, V. Feldman, A. Smith, and K. Talwar. “When is memorization of irrelevant training data necessary for high-accuracy learning?” In: Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing. 2021, pp. 123–132.

The Problem of Memorization

How much a “good” algorithm must reveal about its training data ?

- **Complex relationship between generalization and memorization !**

- Benefits of limited memorization for strong generalization

- D. Russo and J. Zou. How much does your data exploration overfit? Controlling bias via information usage. 2015. arXiv: 1511.05219.
 - R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. “Learners that Use Little Information”. In: Algorithmic Learning Theory. 2018, pp. 25–55.

- But, modern overparametrized DNNs often achieve high accuracy despite memorizing a significant part of the training data !

- N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, “Quantifying memorization across neural language models”’. arXiv preprint arXiv:2202.07646 (2022).
 - V. Feldman and C. Zhang. “What neural networks memorize and why: Discovering the long tail via influence estimation”. Advances in Neural Information Processing Systems 33 (2020), pp. 2881–2891.

- Memorization might even be *necessary* for generalization !

- V. Feldman. “Does learning require memorization ? a short tale about a long tail”’. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. 2020, pp. 954–959.
 - G. Brown, M. Bun, V. Feldman, A. Smith, and K. Talwar. “When is memorization of irrelevant training data necessary for high-accuracy learning?” In: Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing. 2021, pp. 123–132.

- **Informal definition:** a data sample is considered memorized if an adversary can guess correctly if this point appeared in the training set with a high confidence.

Definitions

- **Stochastic Convex Optimization (SCO) problem:** a triple $(\mathcal{W}, \mathcal{Z}, \ell)$, where $\mathcal{W} \in \mathbb{R}^D$ is a **convex set** and $\ell(z, \cdot): \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}$ is a **convex function** for every $z \in \mathcal{Z}$.
- **ϵ -learner for SCO:** $\mathcal{A} = \{\mathcal{A}_n\}_{n \geq 1}$ is called an ϵ -learner algorithm with sample complexity $N: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{N}$, if for every $\delta \in (0, 1]$ and $n \geq N(\epsilon, \delta)$, for every μ , with prob. $1 - \delta$ over S_n , $\mathcal{R}(\mathcal{A}_n(S_n)) - \min_{w \in \mathcal{W}} \mathcal{R}(w) \leq \epsilon$.
- **Recall Game:** Given $\mathcal{A} = \{\mathcal{A}_n\}_{n \geq 1}$, let $\mathcal{Q}: \mathbb{R}^D \times \mathcal{M}_1(\mathcal{Z}) \times \mathcal{Z} \rightarrow \{0, 1\}$ be an adversary for the following game for an $i \in [n]$.
 - Given a **test** data point $Z'_i \sim \mu$ independent of (Z_i, W) , let $Z_{i,0} = Z'_i$ and $Z_{i,1} = Z_i$.
 - Adversary observes Z_{i,K_i} , where $K_i \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(1/2)$.
 - Adversary outputs $\hat{K}_i \triangleq \mathcal{Q}(W, Z_{i,K_i}, \mu)$ as its guess of K_i .
- Consider recall game for n rounds:
 - At each round $i \in [n]$, a pair $(Z_{i,0}, Z_{i,1})$ is considered.
 - The adversary makes two independent guesses: one for $Z_{i,0}$, the other for $Z_{i,1}$.

Memorization and tracing

- **Soundness and recall:** Adversary plays the game in n rounds, twice per round independently of each other, using respectively $(W, Z_{i,0}, \mu)$ and $(W, Z_{i,1}, \mu)$ as input

- **[Test data]** Given $\xi \in [0, 1]$, the adversary is said to be **ξ -sound** if

$$\mathbb{P}\left(\exists i \in [n]: \mathcal{Q}(W, Z_{i,0}, \mu) = 1\right) \leq \xi$$

- **[Training data]** Adversary **certifies the recall of m samples with probability $q \in [0, 1]$** if

$$\mathbb{P}\left(\sum_{i \in [n]} \mathcal{Q}(W, Z_{i,1}, \mu) \geq m\right) \geq q$$

- If both conditions are met, the **adversary (m, q, ξ) -traces the data.**
- **Good adversary $\Rightarrow \xi$ small, m large, and q non-negligible.**
 - A “dummy adversary” can (m, q, ξ) -trace the data if $m = o(n)$ or if $\xi \geq q$.

Memorization for $\mathcal{P}_{cvx}^{(D)}$ problem instances

Convex-Lipschitz-Bounded (CLB) problem: a SCO problem, where $\forall w \in \mathcal{W}$, $\|w\| \leq R$ and the loss function is L -Lipschitz. This class of problems is denoted by $\mathcal{C}_{L,R}$

Problem instance $\mathcal{P}_{cvx}^{(D)} \in \mathcal{C}_{1,1}$ with $\mathcal{Z}, \mathcal{W} \subseteq \mathcal{B}_D(1)$ and $\ell_c(z, w) = -\langle w, z \rangle$ [L23] [ADHLR24]

Memorization for ε -learners of $\mathcal{P}_{cvx}^{(D)}$ [ADHLR24]

Fix **arbitrary** $\xi \in (0, 1]$ and let $\mathcal{Z} = \{\pm 1/\sqrt{D}\}^D$.

Given **any ε -learner algorithm** $\mathcal{A}$ with sample complexity $N(\varepsilon, \delta) = \Theta(\log(1/\delta)/\varepsilon^2)$, there exist

→ a **data distribution** $\mu_{\mathcal{P}^*}$,

→ and an adversary,

such that for $n = N(\varepsilon, \delta)$ and **$D = \Omega(n^4 \log(n/\xi))$** ,

Adversary $(\Omega(n), 1/3, \xi)$ -traces the data

[ADHLR24] I. Attias, G.-K. Dziugaite, M. Haghifam, R. Livni and D. Roy "Information complexity of stochastic convex optimization: Applications to generalization, memorization, and tracing," ICML 2024.

Conditional mutual information (CMI) framework

- $\tilde{\mathbf{S}} \in \mathcal{Z}^{n \times 2}$: a super-sample composed of $2n$ data-points $Z_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mu$, where $j \in \{0, 1\}$ and $i \in [n]$.
- $\mathbf{J} = (J_1, \dots, J_n) \in \{0, 1\}^n$: membership vector, where $J_i \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$
 - $\tilde{\mathbf{S}}_{\mathbf{J}} = \{Z_{1,J_1}, Z_{2,J_2}, \dots, Z_{n,J_n}\}$: plays the role of the training dataset $\mathbf{S}_{\mathbf{n}}$,
 - $\tilde{\mathbf{S}}_{\mathbf{J}^c} = \tilde{\mathbf{S}} \setminus \tilde{\mathbf{S}}_{\mathbf{J}}$: plays the role of a test dataset $\mathbf{S}'_{\mathbf{n}}$
 - $\tilde{\mathbf{S}}$: a shuffled version of the union of the two.
- [SZ20] CMI of an algorithm $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{W}$ with respect to μ :

$$\text{CMI}(\mu, \mathcal{A}) \triangleq \mathbb{I}(\mathcal{A}(\tilde{\mathbf{S}}_{\mathbf{J}}); \mathbf{J} | \tilde{\mathbf{S}})$$

- [SZ20] Assume $\ell(z, w) \in [0, 1]$. Then,

$$\text{gen}(\mu, \mathcal{A}) \leq \sqrt{\frac{2 \text{CMI}(\mu, \mathcal{A})}{n}}$$

[SZ20] Steinke & Zakynthinou. "Reasoning about generalization via conditional mutual information," COLT 2020.

Limitation of CMI for $\mathcal{P}_{cvx}^{(D)}$

CMI bound for $\mathcal{P}_{cvx}^{(D)}$ [ADHLR24]

Consider **any ε -learner algorithm**¹ $\mathcal{A} = \{\mathcal{A}_n\}_{n \in \mathbb{N}}$ for $\mathcal{P}_{cvx}^{(D)}$ with sample complexity $N(\cdot, \cdot)$.

- i. For $n \geq N(\varepsilon, \delta)$ and $\mathbf{D} = \Omega(\mathbf{n}^4 \log(\mathbf{n}))$, there **exists** $\mathcal{Z}$ and a **data distribution** μ^* s.t.

$$\text{CMI}(\mu^*, \mathcal{A}_n) = \Omega\left(\frac{1}{\varepsilon^2}\right).$$

- ii. For optimal sample complexity $N(\varepsilon, \delta) = \Theta\left(\frac{1}{\varepsilon^2}\right)$, the CMI generalization bound equals

$$\text{CMI bound} = LR\sqrt{8\text{CMI}(\mu^*, \mathcal{A}_n)/N(\varepsilon, \delta)} = \Theta(1).$$

¹ **ε -learner for SCO:** $\mathcal{A} = \{\mathcal{A}_n\}_{n \geq 1}$ is called an ε -learner algorithm with sample complexity $N : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{N}$, if for every $\delta \in (0, 1]$ and $n \geq N(\varepsilon, \delta)$, for every μ , with probab. $1 - \delta$ over S_n , $\mathcal{R}(\mathcal{A}_n(S_n)) - \min_{w \in \mathcal{W}} \mathcal{R}(w) \leq \varepsilon$.

Stochastic projection and lossy quantization

- Our new bounds involve two main ingredients, **stochastic projection** and **lossy quantization**.
- **Stochastic projection:**
 - Let $\Theta \in \mathbb{R}^{D \times d'}$ be a random matrix, distributed $\sim P_\Theta$ independently of $\tilde{S}$.
 - Consider the hypothesis $W \in \mathcal{W} \subseteq \mathbb{R}^D$ which lies in a D -dimensional space.
 - Instead of W , we consider its *projection* $\Theta^\top W \in \mathbb{R}^{d'}$ onto a smaller d' -dimensional space.
 - $d' \ll D$.

- **Lossy quantization:**

- The lossy quantization algorithm is a stochastic map $\tilde{\mathcal{A}}: \mathbb{R}^{d'} \rightarrow \hat{\mathcal{W}}$ that maps $\Theta^\top W$ to a “quantized” hypothesis $\hat{W} \in \hat{\mathcal{W}} \subseteq \mathbb{R}^{d'}$, chosen such that

$$E_{P_{S_n, W} P_\Theta P_{\hat{W} | \Theta^\top W}} [\text{gen}(S_n, W) - \text{gen}(S_n, \Theta \hat{W})] \leq \epsilon.$$

- The stochastic map $\tilde{\mathcal{A}}$ induces $P_{\hat{W} | \Theta^\top W}$.

Stochastic projection and lossy quantization

- **Overall ϵ -lossy compression:** Let $\epsilon \in \mathbb{R}$. The overall ϵ -lossy compression algorithm $\hat{\mathcal{A}}: \mathcal{Z}^n \times \mathbb{R}^{D \times d'} \rightarrow \hat{\mathcal{W}}$, is composed of **projection** and **lossy quantization**:

$$\hat{\mathcal{A}}(S_n, \Theta) \triangleq \tilde{\mathcal{A}}\left(\Theta^\top \mathcal{A}(S_n)\right) = \hat{W} \in \mathbb{R}^{d'},$$

and satisfies

$$\text{Distortion} \triangleq \mathbb{E}_{P_{S_n, W} P_{\Theta} P_{\hat{W} | \Theta^\top W}} \left[\text{gen}(S_n, W) - \text{gen}(S_n, \Theta \hat{W}) \right] \leq \epsilon.$$

- **Disintegrated CMI:** For a super-sample $\tilde{S}$ and a stochastic projection matrix Θ :

$$\text{CMI}^{\Theta}(\tilde{S}, \hat{\mathcal{A}}) \triangleq \mathcal{I}^{\tilde{S}, \Theta}(\hat{\mathcal{A}}(\tilde{S}_{\mathbf{J}}, \Theta); \mathbf{J})$$

where $\mathcal{I}^{\tilde{S}, \Theta}(\hat{\mathcal{A}}(\tilde{S}_{\mathbf{J}}, \Theta); \mathbf{J})$ is the CMI given an instance of $\tilde{S}$ and Θ , computed $\sim P_{\mathbf{J}} \otimes P_{W | \tilde{S}_{\mathbf{J}}} \otimes P_{\hat{W} | \Theta^\top W}$, with $P_{\mathbf{J}} = \text{Bern}(1/2)^{\otimes n}$.

[SKZ, NeurIPS 2025] M. Sefidgaran, K. Nadjahi and A. Zaidi "Tighter CMI-Based Generalization Bounds via Stochastic Projection and Quantization" NeurIPS 2025.

Projected-quantized CMI bound

Projected-quantized CMI bound

For every $\epsilon \in \mathbb{R}$, every $d' \in \mathbb{N}$, and every *projected model quantization set* $\hat{\mathcal{W}} \subseteq \mathbb{R}^{d'}$,

$$\text{gen}(\mu, \mathcal{A}) \leq \inf_{P_{\hat{\mathcal{W}}|\Theta^\top W}} \inf_{P_\Theta} \mathbb{E}_{P_{\tilde{\mathbf{S}}} P_\Theta} \left[\sqrt{\frac{2\Delta\ell_{\hat{\mathcal{W}}}(\tilde{\mathbf{S}}, \Theta)}{n}} \text{CMI}^\Theta(\tilde{\mathbf{S}}, \hat{\mathcal{A}}) \right] + \epsilon,$$

where $\hat{W} \in \hat{\mathcal{W}}$, $\Theta \in \mathbb{R}^{D \times d'}$, the infima are over all $P_{\hat{\mathcal{W}}|\Theta^\top W}$ and P_Θ such that:

$$\text{Distortion} := \mathbb{E}_{P_{S_n, W} P_\Theta P_{\hat{\mathcal{W}}|\Theta^\top W}} [\text{gen}(S_n, W) - \text{gen}(S_n, \Theta \hat{W})] \leq \epsilon,$$

and

$$\Delta\ell_{\hat{\mathcal{W}}}(\tilde{\mathbf{S}}, \Theta) := \mathbb{E}_{P_{W|\tilde{\mathbf{S}}} P_{\hat{\mathcal{W}}|\Theta^\top W}} \left[\frac{1}{n} \sum_{i \in [n]} (\ell(Z_{i,0}, \Theta \hat{W}) - \ell(Z_{i,1}, \Theta \hat{W}))^2 \right].$$

[SKZ, NeurIPS 2025] M. Sefidgaran, K. Nadjahi and A. Zaidi "Tighter CMI-Based Generalization Bounds via Stochastic Projection and Quantization" NeurIPS 2025.

Projected-quantized CMI bound for $\mathcal{P}_{cvx}^{(D)}$

Projected-Quantized CMI bound for $\mathcal{P}_{cvx}^{(D)}$

For every $\mathcal{A}: \mathcal{Z}^n \rightarrow \mathcal{W}$ of the instance $\mathcal{P}_{cvx}^{(D)}$,

$$\text{gen}(\mu, \mathcal{A}) \leq \text{Projected-Quantized CMI bound} = \frac{8}{\sqrt{n}}.$$

In particular, setting $N(\varepsilon, \delta) = \Theta\left(\frac{1}{\varepsilon^2}\right)$ for ε -learner algorithms we get

$$\text{gen}(\mu, \mathcal{A}) = \mathcal{O}(\varepsilon).$$

- Impossibility result of [ADHLR24] is obtained for an ε -learner and a specific choice of $\mathcal{Z}$ and μ^* .
- The above result holds for **any learning algorithm, any $\mathcal{Z} \subseteq \mathcal{B}_D(1)$, and any μ .**

[SKZ, NeurIPS 2025] M. Sefidgaran, K. Nadjahi and A. Zaidi "Tighter CMI-Based Generalization Bounds via Stochastic Projection and Quantization" NeurIPS 2025.

Memorization for $\mathcal{P}_{cvx}^{(D)}$ problem instances

Untraceability of the projected-quantized model (1/2)

Fix **arbitrary** $r > 0$ and **arbitrary** $\mathcal{Z} \subseteq \mathcal{B}_D(1)$.

For **any learning algorithm** $\mathcal{A}: \mathcal{Z}^n \rightarrow \mathbb{R}^D$, there exists a **projected-quantized algorithm** $\mathcal{A}^*: \mathcal{Z}^n \rightarrow \mathbb{R}^D$, defined as

$$\mathcal{A}^*(S_n) \triangleq \Theta \tilde{\mathcal{A}}(\Theta^\top \mathcal{A}(S_n)) = \Theta \hat{W},$$

where $\Theta \in \mathbb{R}^{D \times d'}$, $\Theta \sim P_\Theta$ independent of (S_n, W) for $d' = 500r \log(n)$, such that for **any data distribution** μ , the following conditions are met simultaneously:

- i. Generalization error of the auxiliary model $\Theta \hat{W}$ satisfies

$$\left| \mathbb{E}_{P_{S_n, W} P_\Theta P_{\hat{W} | \Theta^\top W}} [\text{gen}(S_n, W) - \text{gen}(S_n, \Theta \hat{W})] \right| = \mathcal{O}(n^{-r}),$$

[SKZ, NeurIPS 2025] M. Sefidgaran, K. Nadjahi and A. Zaidi "Tighter CMI-Based Generalization Bounds via Stochastic Projection and Quantization" NeurIPS 2025.

Memorization for $\mathcal{P}_{cvx}^{(D)}$ problem instances

Untraceability of the projected-quantized model (2/2)

- ii. If there exists an adversary which, by **having access to both Θ and $\hat{W}$** (hence, $\Theta\hat{W}$), (m, q, ξ) -traces the data, then
- $m = o(n)$ **or** $\xi \geq q$ (**not better than a dummy adversary**)
 - if $m = \Omega(n)$ and $\xi = \Omega(1)$ (**it fails on a non-negligible portion of test samples**)

$$\mathbb{P}\left(\sum_{i \in [n]} \mathcal{L}(\Theta\hat{W}, Z_{i,0}, \mu) \geq \Omega(n)\right) \geq \Omega(1)$$

- Proof idea, based on **Fano's inequality for approximate recovery**,
 - Constructing an estimator of the index set $\mathbf{J}$ based on the adversary's guesses,
 - Showing if this estimator can correctly recover a fraction $c > \frac{1}{2}$ of $\mathbf{J}$ indices, then $\text{CMI}^\Theta(\mu, \mathcal{A}^*) = \Theta(n)$.
- Similar results exist for the following cases:
 - if **population risk** closeness is considered instead of generalization error
 - for **deterministic** Θ (at the expense of the compressed algorithm being dependent on data distribution)

Discussion

- Consider $D = \Omega(n^4 \log(n/\xi))$ and **any ϵ -learner algorithm $\mathcal{A}$** with output W for $\mathcal{P}_{cvx}^{(D)}$ problem instances
 - [ADHLR24] shows that there **exists a data distribution** for which $\mathcal{A}$ **must memorize** a large fraction of the training/test data.
 - Our results show that the auxiliary model $\Theta \hat{W}$
 - (i) **does not memorize** the training/test data for **any data distribution**,
 - (ii) on average over Θ , **generalization errors** for models $\Theta \hat{W}$ and W are **arbitrarily close**.
- **Contradiction?**
 - **No!** $\Theta \hat{W}$ does not satisfy the bounded conditions required in [ADHLR24].
 - In particular, for any w , **while** $\mathbb{E}_{\hat{W}, \Theta} [\Theta \hat{W}] \approx w$, **but** $\mathbb{E}_{\hat{W}, \Theta} [\|\Theta \hat{W}\|^2] = \Omega(D/d') = \Omega(n^3)$.

[ADHLR24] Attias et al. "Information complexity of stochastic convex optimization: Applications to generalization, memorization, and tracing," ICML 2024.

Construction of projected-quantized model via Johnson-Lindenstrauss transform

- Fix some constants $c_w \in \left[1, \sqrt{\frac{5}{4}}\right)$, $\nu \in (0, 1]$, and $d' \in \mathbb{N}^*$.
- Let $\Theta \in \mathbb{R}^{D \times d'}$, with elements $\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d'})$.
- Given Θ and $W = \mathcal{A}(S_n)$, in the scheme $\mathbf{JL}(d', c_w, \nu)$, let

$$U := \begin{cases} \Theta^\top W, & \text{if } \|\Theta^\top W\| \leq c_w, \\ \mathbf{0}_{d'}, & \text{otherwise.} \end{cases}$$

- Let $\hat{W} \in \hat{\mathcal{W}} = \mathcal{B}_{d'}(c_w + \nu)$ be defined as

$$\hat{W} = U + V_\nu \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}\left(\mathcal{B}_{d'}(U, \nu)\right),$$

with V_ν : independent random variable sampled uniformly on $\mathcal{B}_{d'}(\nu)$

- **This defines P_Θ and $P_{\hat{W}|\Theta^\top W}$ for a $\mathbf{JL}(d', c_w, \nu)$ scheme.**

Properties of $JL(d', c_w, \nu)$

- Disintegrated CMI:

$$\text{CMI}^\Theta(\tilde{\mathbf{S}}, \hat{\mathcal{A}}) \leq d' \log\left(\frac{c_w + \nu}{\nu}\right)$$

- Loss difference:

$$\mathbb{E}_{P_{\tilde{\mathbf{S}}} P_{\Theta}} [\Delta \ell_{\hat{w}}(\tilde{\mathbf{S}}, \Theta)] \leq 4(c_w + \nu)^2$$

- Distortion:

$$\text{Distortion} \leq \frac{3}{(\sqrt{n} \text{ or } 1)} e^{-\frac{0.21}{4} d' (c_w^2 - 1)^2}$$

- Projected-quantized CMI bound:

$$\text{gen}(\mu, \mathcal{A}) \leq \mathbb{E}_{P_{\tilde{\mathbf{S}}} P_{\Theta}} \left[\sqrt{\frac{2\Delta \ell_{\hat{w}}(\tilde{\mathbf{S}}, \Theta)}{n}} \text{CMI}^\Theta(\tilde{\mathbf{S}}, \hat{\mathcal{A}}) \right] + \text{Distortion}$$

- In our proofs, $d' \in \{1, \mathcal{O}(n^r), \mathcal{O}(\log(n))\}$ for some $r \in \mathbb{R}_+$

Properties of $JL(d', c_w, \nu)$

- Hence, for the **counterexample of [ADHLR24]**,
 - **CMI of the original model blows up as ε increases: $CMI = \Omega(1/\varepsilon^2) = \Omega(N(\varepsilon, \delta))$**
 - **(Disintegrated) CMI of the projected quantized model is negligible: $\mathcal{O}(d')$,**
 - **Generalization-wise; two models are very close: having difference of $\mathcal{O}\left(\frac{e^{-\alpha d'}}{\sqrt{n}}\right)$.**
- **Similar results hold for**
 - counterexample of [ADHLR24] for Convex set-Strongly Convex-Lipschitz (CSL) subclass,
 - counterexample of [L23],
 - generalized linear stochastic optimization problems.

[L23] Livni. Information theoretic lower bounds for information theoretic upper bounds," NeurIPS 2023.

[ADHLR24] Attias et al. "Information complexity of stochastic convex optimization: Applications to generalization, memorization, and tracing," ICML 2024.

Implications and Summary

- **Existence of sample-compression schemes**

- If $\mathcal{A}_n = \psi(\phi(S))$ is a sample-compression scheme of size k then it must be $\text{CMI}(\mathcal{A}_n) \leq k \log 2n$ [Steinke-Zakynthinou'20].
- From Attias et al., there are **no** sample-compression schemes for $\mathcal{P}_{cvx}$ (CMI blows up faster than n)!
- A path to constructing sample-compression schemes when the definition is extended to involve approximate reconstruction (in terms of induced generalization error) instead of the strict $\mathcal{A}_n(\cdot) = \psi(\phi(\cdot))$ of Littlestone and Warmuth.

- **Fingerprinting codes and privacy attacks**

- For designing privacy attacks on mean estimators that expose a fraction of the training data, a well-designed adversary can guess membership of the training samples from the output of every algorithm that estimates mean with high precision [Dwork et al., 2015].
- Our results might be useful to construct differentially private codes that prevent such fingerprinting type attacks (by injecting noise in a suitable smaller subspace).

- **Summary:**

- We introduced **stochastic projection** together with **lossy quantization** within the CMI framework, and used them to establish a new **lossy-CMI-based** generalization bound.
- Our new bound attains the **optimal order-wise rate** for counterexamples where the **CMI bound fails**.
- For SCO problems for which **“good” learning algorithm must memorize under a given data distribution**, we showed that there **exists a closely projected-quantized model that does not memorize under any data distribution**.

References

- [M98] McAllester. "Some PAC-Bayesian theorems," COLT 1998.
- [XR17] Xu and Raginsky. "Information-theoretic analysis of generalization capability of learning algorithms," NeurIPS 2017.
- [SZ20] Steinke and Zakynthinou. "Reasoning about generalization via conditional mutual information," COLT 2020.
- [NDR20] Negrea, Dziugaite, and Roy. "In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors," ICML 2020.
- [GKL24] Grønlund, Kamma, and Larsen. "Near-tight margin-based generalization bounds for support vector machines," ICML 2024.
- [SGRS22] Sefidgaran, Gohari, Richard, and Şimşekli. "Rate-distortion theoretic generalization bounds for stochastic learning algorithms," COLT 2022.
- [SCZ22] Sefidgaran, Chor, and Zaidi. "Rate-distortion theoretic bounds on generalization error for distributed learning," NeurIPS 2022.
- [HRTSRD23] Haghifam, Rodríguez-Gálvez, Thobaben, Skoglund, Roy, and Dziugaite. "Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization," ALT 2023.
- [L23] Livni. "Information theoretic lower bounds for information theoretic upper bounds," NeurIPS 2023.
- [SZ24] Sefidgaran and Zaidi. "Data-dependent generalization bounds via variable-size compressibility," IEEE Transactions on Information Theory 2024.
- [ADHLR24] Attias, Dziugaite, Haghifam, Livni, and Roy. "Information complexity of stochastic convex optimization: Applications to generalization, memorization, and tracing," ICML 2024.
- [KGBS24] Nadjahi, Greenewald, Gabrielsson, and Solomon. "Slicing mutual information generalization bounds for neural networks," ICML 2024.

Thank you!