

Tighter CMI-based Generalization Bounds via Stochastic Projection and Quantization

Milad Sefidgaran¹

Kimia Nadjahi²

Abdellatif Zaidi^{1,3}

¹ Paris Research Center, Huawei Technologies France

² CNRS, ENS Paris, France

³ Université Gustave Eiffel, France



Neural Information Processing Systems (NeurIPS), December 2025

Outline

Problem setup and motivation

Lossy algorithm compression

Projected-quantized CMI bound

Resolving recently raised limitations of classic CMI bounds

Memorization

Implications and Conclusion

Milad Sefidgaran

Problem Setup
●○○○○

Lossy compression
○○○○○

Projected-Quantized CMI
○○○○

IT Counterexamples
○○○○○○○

Memorization
○○○○○○○

Conclusion
○○○

- Data $Z \in \mathcal{Z}$ distributed according to an unknown distribution μ
- Training dataset $S_n = \{Z_1, \dots, Z_n\} \sim P_{S_n} = \mu^{\otimes n}$
- **Randomized algorithm** $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{W}$:
 - takes S_n as input and chooses a hypothesis $\mathcal{A}(S_n) = W \in \mathcal{W}$
 - induces a conditional distribution $P_{W|S_n}$
- **Loss function** $\ell : \mathcal{Z} \times \mathcal{W} \rightarrow \mathbb{R}$
- **Population risk**: $\mathcal{R}(w) \triangleq \mathbb{E}_{Z \sim \mu}[\ell(Z, w)]$ and **Empirical risk**: $\widehat{\mathcal{R}}(s_n, w) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(z_i, w)$

$$\text{Generalization error} \quad \text{gen}(s_n, w) \triangleq \mathcal{R}(w) - \widehat{\mathcal{R}}(s_n, w)$$

Some information-theoretic generalization bounds

- [M98] Fix some prior Q_W and assume $\ell(z, w) \in [0, 1]$. Then, with probability $1 - \delta$ over $S_n \sim \mu^{\otimes n}$,

$$\mathbb{E}_{W \sim P_{W|S_n}} [\text{gen}(S_n, W)] \leq \sqrt{\frac{D_{KL}(P_{W|S_n} \| Q_W) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}$$

- [XR17] Assume $\ell(z, w) \in [0, 1]$. Then,

$$\mathbb{E}_{S_n, W \sim P_{S_n, W}} [\text{gen}(S_n, W)] \triangleq \text{gen}(\mu, \mathcal{A}) \leq \sqrt{\frac{I(S_n; W)}{2n}}$$

[M98] McAllester. “Some PAC-Bayesian theorems,” COLT 1998.

[XR17] Xu & Raginsky. “Information-theoretic analysis of generalization capability of learning algorithms” NeurIPS 2017.

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 $\mathbf{S}_n$,
 - $\tilde{\mathbf{S}}_{\mathbf{J}^c} = \tilde{\mathbf{S}} \setminus \tilde{\mathbf{S}}_{\mathbf{J}}$: plays the role of a test dataset $\mathbf{S}'_n$
 - $\tilde{\mathbf{S}}$: 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.

Motivation: raised information-theoretic limitations

- Several papers have studied the limitations of information-theoretic generalization bounds.
- In particular, [HRTSRD23] [L23] [ADHLR24] have provided **counterexamples** where:
 - many **information-theoretic (IT)** generalization bounds become **vacuous**.
 - any **“good” learning algorithm ‘must’ memorize** the training data for a data distribution!
- **Two main questions** in our work:

1. Do presented counterexamples reveal intrinsic limitations of IT approaches?
2. Is memorization inevitable for effective learning?

[HRTSRD23] Haghifam et al. “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.

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

Outline

Problem setup and motivation

Lossy algorithm compression

Projected-quantized CMI bound

Resolving recently raised limitations of classic CMI bounds

Memorization

Implications and Conclusion

Milad Sefidgaran

Problem Setup
○○○○○

Lossy compression
●○○○○

Projected-Quantized CMI
○○○○

IT Counterexamples
○○○○○○○

Memorization
○○○○○○○

Conclusion
○○○

A closer look into information-theoretic bounds

- Assume $\ell(z, w) \in [0, 1]$ and $|\mathcal{W}| < \infty$. Then, $\text{gen}(\mu, \mathcal{A}) \leq \sqrt{\frac{\log(|\mathcal{W}|)}{2n}}$, by maximal inequality.
- [SGRS22] [SZ24] Aforementioned information-theoretic bounds can be obtained by first **applying ‘lossless block-coding compression’** and then **invoking the above bound**.

[SGRS22] Sefidgaran et al. “Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms,” COLT 2022.

[SZ24] Sefidgaran & Zaidi. “Data-dependent generalization bounds via variable-size compressibility,” IEEE Transactions on Information Theory 2024.

A closer look into information-theoretic bounds

- Assume $\ell(z, w) \in [0, 1]$ and $|\mathcal{W}| < \infty$. Then, $\text{gen}(\mu, \mathcal{A}) \leq \sqrt{\frac{\log(|\mathcal{W}|)}{2n}}$, by maximal inequality.
- [SGRS22] [SZ24] Aforementioned information-theoretic bounds can be obtained by first **applying ‘lossless block-coding compression’** and then **invoking the above bound**.
- Consider m independent training datasets $S^m = (S_n(1), \dots, S_n(m)) \in \mathcal{Z}^{nm}$ and the corresponding picked hypotheses $W^m = (W(1), \dots, W(m)) \in \mathcal{W}^m$, where $W(j) = \mathcal{A}(S_n(j))$.
- **By covering lemma**, $\exists$ **hypothesis book** $\mathcal{H}_m \subseteq \mathcal{W}^m$, $\forall \hat{\mathbf{w}} \in \mathcal{H}_m: \hat{\mathbf{w}} = (\hat{w}(1), \dots, \hat{w}(m)) \in \mathcal{W}^m$, s.t.
 - With probability $P_m \xrightarrow{m \rightarrow \infty} 1$, for (S^m, W^m) , $\exists \hat{\mathbf{w}}^* \in \mathcal{H}_m$ such that $\hat{p}_{(S^m, W^m)} = \hat{p}_{(S^m, \hat{\mathbf{w}}^*)}$,
 - $|\mathcal{H}_m| \lesssim e^{ml(S_n; W)}$.

[SGRS22] Sefidgaran et al. “Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms,” COLT 2022.

[SZ24] Sefidgaran & Zaidi. “Data-dependent generalization bounds via variable-size compressibility,” IEEE Transactions on Information Theory 2024.

A closer look into information-theoretic bounds

- By letting $m \rightarrow \infty$,

$$\begin{aligned}\text{gen}(\mu, \mathcal{A}) &= \frac{1}{m} \mathbb{E}_{S^m, W^m} \left[\sum_{j \in [m]} \text{gen}(S_n(j), W(j)) \right] \\ &\xrightarrow{\text{green}} \frac{1}{m} \mathbb{E}_{S^m, \hat{W}^*} \left[\sum_{j \in [m]} \text{gen}(S_n(j), \hat{W}^*(j)) \right] \quad (\text{since } \hat{p}_{(S^m, W^m)} = \hat{p}_{(S^m, \hat{W}^*)}) \\ &\leq \sqrt{\frac{\log(e^{\text{red } m I(S_n; W)})}{2n \text{red } m}} = \sqrt{\frac{I(S_n; W)}{2n}}.\end{aligned}$$

- $\Rightarrow I(S_n; W)$ is an **upper bound** on **lossless compressibility** level of $\mathcal{A}$. [SGRS22]
- $\Rightarrow \text{CMI}(\mu, \mathcal{A})$ is an **upper bound** on **lossless compressibility** level of $\mathcal{A}$, given $\tilde{\mathbf{S}}$. [SGRS22]
- $\Rightarrow D_{KL}(P_{W|S_n} \| Q_W)$ is an **upper bound** on **lossless variable-size compressibility** level of $\mathcal{A}$, given S_n . [SZ24]

[SGRS22] Sefidgaran et al. “Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms,” COLT 2022.

[SZ24] Sefidgaran & Zaidi. “Data-dependent generalization bounds via variable-size compressibility,” IEEE Transactions on Information Theory 2024.

How to understand raised limitations?

- By **source-coding** and **coordination** results, **lossless compression** (coverings) of
 - **continuous sources** or mappings requires **infinite rate!**
 - **high-dimensional sources** or mappings **often provides negligible reduction!**
- Naturally, for effective compression, one needs to consider **lossy compression**:
 $\Rightarrow$ to find $\hat{\mathbf{w}}^* \in \mathcal{H}_m$ such that $\sum_{j \in [m]} \text{gen}(S_n(j), W(j)) \approx \sum_{j \in [m]} \text{gen}(S_n(j), \hat{\mathbf{w}}^*(j))$.
- In this work, we follow [NDR20] [SGRS22] to use the **Rate-Distortion theoretic** approach by finding a suitable “**surrogate**” or “**compressed**” algorithm.
 - Studies found that high-dimensional trained models often reside in a low-dimensional subspace.
 - Inspired by this, and following [GKL20] [SCZ22] [KGBS24], we build the lossy algorithm by **stochastic projection** and **quantization**.

[NDR20] Negrea et al. “In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors,” ICML 2020.

[GKL20] Grönlund et al. “Near-tight margin-based generalization bounds for support vector machines,” ICML 2024.

[SGRS22] Sefidgaran et al. “Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms,” COLT 2022.

[SCZ22] Sefidgaran et al. “Rate-distortion theoretic bounds on generalization error for distributed learning” NeurIPS 2022.

[KGBS24] Nadjahi et al. “Slicing mutual information generalization bounds for neural networks,” ICML 2024.

- We introduce **stochastic projection** and **lossy quantization** within the CMI framework and use them to establish a new **lossy–CMI–based** generalization bound.
- We show that the new bound attains the **optimal order-wise rate** for counterexamples where the **CMI bound fails**.
- For counterexamples in which any **“good” learning algorithm must memorize under a given data distribution**, we show that there **exists a ‘close’ projected–quantized model that does not memorize under any data distribution**.

Outline

Problem setup and motivation

Lossy algorithm compression

Projected-quantized CMI bound

Resolving recently raised limitations of classic CMI bounds

Memorization

Implications and Conclusion

Milad Sefidgaran

Problem Setup
○○○○○

Lossy compression
○○○○○

Projected-Quantized CMI
●○○○

IT Counterexamples
○○○○○○○

Memorization
○○○○○○○

Conclusion
○○○

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 $\tilde{\mathbf{S}}$.
 - 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 $\hat{W} \in \hat{\mathcal{W}} \subseteq \mathbb{R}^{d'}$.
 - 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**:

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

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 $\tilde{\mathbf{S}}$ and a stochastic projection matrix Θ :

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

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

Projected-quantized CMI bound

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

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

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

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

and

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

Outline

Problem setup and motivation

Lossy algorithm compression

Projected-quantized CMI bound

Resolving recently raised limitations of classic CMI bounds

Memorization

Implications and Conclusion

Milad Sefidgaran

Problem Setup
○○○○○

Lossy compression
○○○○○

Projected-Quantized CMI
○○○○

IT Counterexamples
●○○○○○

Memorization
○○○○○○○

Conclusion
○○○

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 $\ell(z, \cdot): \mathcal{W} \rightarrow \mathbb{R}$ is a **convex function** for every $z \in \mathcal{Z}$.
- **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 $\mathcal{C}_{L,R}$
- CMI generalization bound [HRTSRD23] for $\mathcal{C}_{L,R}$:

$$\text{gen}(\mu, \mathcal{A}) \leq LR \sqrt{\frac{8}{n} \text{CMI}(\mu, \mathcal{A})}$$

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

[HRTSRD23] Haghifam et al. “Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization,” ALT 2023.

- **Stochastic Convex Optimization (SCO) problem:** a triple $(\mathcal{W}, \mathcal{Z}, \ell)$, where $\mathcal{W} \in \mathbb{R}^D$ is a **convex set** and $\ell(z, \cdot): \mathcal{W} \rightarrow \mathbb{R}$ is a **convex function** for every $z \in \mathcal{Z}$.
- **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 $\mathcal{C}_{L,R}$
- CMI generalization bound [HRTSRD23] for $\mathcal{C}_{L,R}$:

$$\text{gen}(\mu, \mathcal{A}) \leq LR \sqrt{\frac{8}{n} \text{CMI}(\mu, \mathcal{A})}$$

- **Problem instance** $\mathcal{P}_{cvx}^{(D)} \in \mathcal{C}_{1,1}$ [L23] [ADHLR24]: Let $\mathcal{Z}, \mathcal{W} \subseteq \mathcal{B}_D(1)$ and

$$\ell_c(z, w) = -\langle w, z \rangle$$

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

[HRTSRD23] Haghifam et al. “Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization,” ALT 2023.

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

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

- i. For $n \geq N(\varepsilon, \delta)$ and $\mathbf{D} = \Omega(n^4 \log(n))$, 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 , $\mathcal{R}(\mathcal{A}_n(S_n)) - \min_{w \in \mathcal{W}} \mathcal{R}(w) \leq \varepsilon$.

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

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 .**

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 $\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{d'}\right)$.
- Given Θ and $W = \mathcal{A}(S_n)$, in the scheme **JL**(d', c_w, ν), 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 **JL**(d', c_w, ν) scheme.**

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

- 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 $r \in \mathbb{R}_+$

- Hence, for the **counterexample of [ADHLR24]**,
 - **CMI of the original model blows up as ε increases: $\text{CMI} = \Omega(1/\varepsilon^2) = \Omega(N(\epsilon, \delta))$**
 - **(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.

Outline

Problem setup and motivation

Lossy algorithm compression

Projected-quantized CMI bound

Resolving recently raised limitations of classic CMI bounds

Memorization

Implications and Conclusion

Milad Sefidgaran

Problem Setup
○○○○○

Lossy compression
○○○○○

Projected-Quantized CMI
○○○○

IT Counterexamples
○○○○○○○

Memorization
●○○○○○

Conclusion
○○○

Memorization and recall game

- **Recall Game** [ADHLR24]: Given $\mathcal{A} = \{\mathcal{A}_n\}_{n \geq 1}$, let $Q: \mathbb{R}^D \times \mathcal{M}_1(\mathcal{Z}) \times \mathcal{Z} \rightarrow \{0, 1\}$ 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 $\hat{K}_i \triangleq Q(W, Z_{i,K_i}, \mu)$ 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 $Z_{i,1}$.

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

Memorization and tracing

- **Soundness and recall** [ADHLR24]: 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 $\xi \geq q$.

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

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** μ_{p^*} ,

→ and an adversary,

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

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

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

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

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

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
 - a. $m = o(n)$ **or** $\xi \geq q$ (**not better than a dummy adversary**)
 - b. if $m = \Omega(n)$ and $q = \Omega(1)$ (**it fails on a non-negligible portion of test samples**)

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

- 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)

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

- Consider $\mathbf{D} = \Omega(n^4 \log(n/\xi))$ and **any ε -learner algorithm $\mathcal{A}$** with output W
- [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.

Problem setup and motivation

Lossy algorithm compression

Projected-quantized CMI bound

Resolving recently raised limitations of classic CMI bounds

Memorization

Implications and Conclusion

- **Implications**

- Differential privacy
- Sample-compression schemes

- **Main Contributions**

- We introduced **stochastic projection** together with **lossy quantization** within the CMI framework, and use them to establish a new **lossy-CMI-based** generalization bound.
- We showed that the new bound attains the **optimal order-wise rate** for counterexamples where the **CMI bound fails**.
- For counterexamples in which any **“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**.
- We established novel generalization bounds for subspace training algorithms using SGLD/SGD.

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 Zakyntinou. “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.