A PAC-BAYESIAN LINK BETWEEN GENERALISATION AND FLAT MINIMA

JOINT WORK WITH PAUL VIALLARD, UMUT SIMSEKLI AND BENJAMIN GUEDJ

Maxime Haddouche

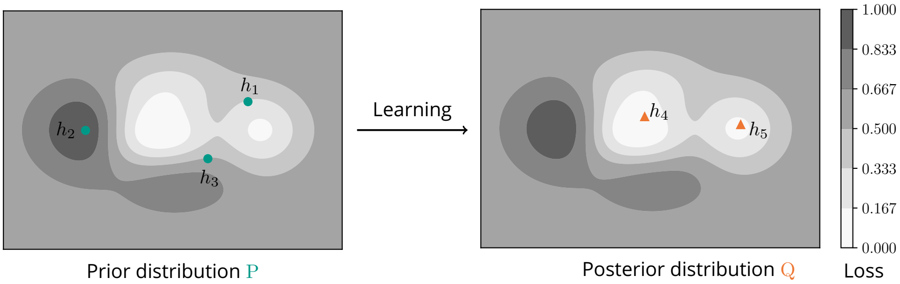
INRIA London
Université de Lille

Session MALIA JdS 2024

(SPECIAL CASE OF) PAC-BAYESIAN LEARNING

PAC-Bayesian learning

Learning a distribution $\mathbb{Q}$ over models from the data and a prior distribution $\mathbb{P}$



Prior distribution $\mathbb{P}$ — Posterior distribution $\mathbb{Q}$

PAC-Bayesian generalisation bounds in a nutshell

With probability at least $1-\delta$

$$\text{performance gap}(\mathbb{Q}) \leq \text{bound}\left(\text{complexity}(\mathbb{Q},\mathbb{P}), \frac{1}{m}, \ln\frac{1}{\delta}\right)$$

SETTING

Notations:

- Predictor/hypothesis $h \in \mathcal{H}$, Data space $\mathcal{Z}$
- Loss $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$, *possibly heavy-tailed*
- m-sized *i.i.d.* learning sample $\mathcal{S} \in \mathcal{Z}^m$, $\mathcal{S} := \{\mathbf{z}_i\}_{i=1}^m \sim \mathcal{D}^{\otimes m}$
- Population risk $\mathsf{R}_{\mathcal{D}}(h) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\, \ell(h, \mathbf{z})$ and empirical risk $\hat{\mathsf{R}}_{\mathcal{S}}(h) = \frac{1}{m}\sum_{i=1}^m \ell(h, \mathbf{z}_i)$
- Expected risks $\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) = \underset{h\sim\mathrm{Q}}{\mathbb{E}}[\mathsf{R}_{\mathcal{D}}(h)], \quad \hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{Q}) = \underset{h\sim\mathrm{Q}}{\mathbb{E}}[\hat{\mathsf{R}}_{\mathcal{S}}(h)]$
- Space of distributions over $\mathcal{H}$: $\mathcal{M}(\mathcal{H})$

Catoni's bound Alquier *et al.* (2016, Theorem 4.1) (σ-subgaussian losses)

For $\lambda > 0$, with probability $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^m$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \ln\frac{1}{\delta}}{\lambda} + \frac{\lambda\sigma^2}{2m}$$

FLAT MINIMUM

What is a flat minimum?

What is a flat minimum?

A minimum such that its neighbourhood nearly minimises the loss.

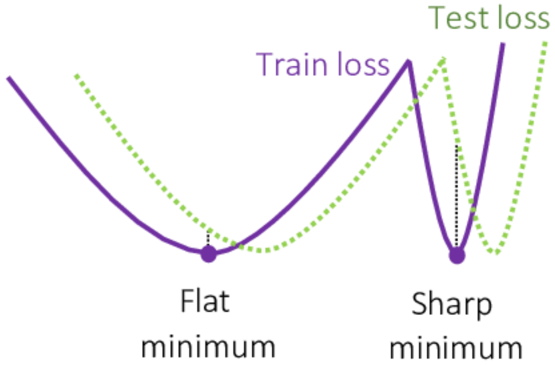


Image from Liebenwein *et al.* (2021).

FLAT MINIMA AND GENERALISATION ARE CORRELATED!

Correlations with generalisation recently emerged:

- Flat minima of $\hat{\mathrm{R}}_{\mathcal{S}}$.
  PAC-Bayes based correlation measure : works for many datasets (Neyshabur *et al.*, 2017; Dziugaite *et al.*, 2020; Jiang *et al.*, 2020)

- Flat minima of the adversarial loss in the context of adversarially robust learning. (Stutz *et al.*, 2021)

- Flat minima implies generalisation for 2-layers nets (Wen *et al.*, 2023).

FLAT MINIMA AND GENERALISATION ARE CORRELATED!

Correlations with generalisation recently emerged:

- Flat minima of $\hat{R}_{\mathcal{S}}$.
  PAC-Bayes based correlation measure : works for many datasets (Neyshabur *et al.*, 2017; Dziugaite *et al.*, 2020; Jiang *et al.*, 2020)

- Flat minima of the adversarial loss in the context of adversarially robust learning. (Stutz *et al.*, 2021)

- Flat minima implies generalisation for 2-layers nets (Wen *et al.*, 2023).

Can we go beyond correlation or 2-layers net and obtain sound generalisation bounds involving directly flat minima?

Notation: for any Q, $\mathrm{H}^1(\mathrm{Q}) := \left\{f \in \mathrm{L}^2(\mathrm{Q}) \cap \mathrm{D}_1(\mathbb{R}^d) \mid \|\nabla f\| \in \mathrm{L}^2(\mathrm{Q})\right\}$

Poincaré

Q is $\texttt{Poinc}(c_P)$ if for all $f \in \mathrm{H}^1(\mathrm{Q})$:

$$\mathrm{Var}(f) \leq c_P(\mathrm{Q}) \underset{h\sim\mathrm{Q}}{\mathbb{E}}\left[\|\nabla f(h)\|^2\right],$$

Log-Sobolev

Q is $\texttt{L-Sob}(c_{LS})$ if for all function $f \in \mathrm{H}^1(\mathrm{Q})$:

$$\underset{h\sim\mathrm{Q}}{\mathbb{E}}\left[f^2(h)\log\left(\frac{f^2(h)}{\mathbb{E}_{h\sim\mathrm{Q}}\left[f^2(h)\right]}\right)\right] \leq c_{LS}(\mathrm{Q}) \underset{h\sim\mathrm{Q}}{\mathbb{E}}\left[\|\nabla f(h)\|^2\right],$$

ESSENTIAL TOOLS: POINCARÉ AND LOG-SOBOLEV INEQUALITIES

Notation: for any $\mathbb{Q}$, $\mathrm{H}^1(\mathbb{Q}) := \left\{ f \in \mathrm{L}^2(\mathbb{Q}) \cap \mathrm{D}_1(\mathbb{R}^d) \mid \|\nabla f\| \in \mathrm{L}^2(\mathbb{Q}) \right\}$

Poincaré

$\mathbb{Q}$ is $\texttt{Poinc}(c_P)$ if for all $f \in \mathrm{H}^1(\mathbb{Q})$:

$$\mathrm{Var}(f) \leq c_P(\mathbb{Q}) \underset{h\sim\mathbb{Q}}{\mathbb{E}} \left[\|\nabla f(h)\|^2 \right],$$

Log-Sobolev

$\mathbb{Q}$ is $\texttt{L-Sob}(c_{LS})$ if for all function $f \in \mathrm{H}^1(\mathbb{Q})$:

$$\underset{h\sim\mathbb{Q}}{\mathbb{E}} \left[f^2(h) \log \left(\frac{f^2(h)}{\mathbb{E}_{h\sim\mathbb{Q}}\left[f^2(h)\right]} \right) \right] \leq c_{LS}(\mathbb{Q}) \underset{h\sim\mathbb{Q}}{\mathbb{E}} \left[\|\nabla f(h)\|^2 \right],$$

Gaussian distributions and Gibbs posteriors are `Poinc` and `L-Sob`!

FAST-RATE GENERALISATION BOUNDS FOR FLAT MINIMA (1)

Notation: $\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z}) := \mathbb{E}_{h\sim \mathrm{Q}}[\ell(h, \mathbf{z})]$

Assumption

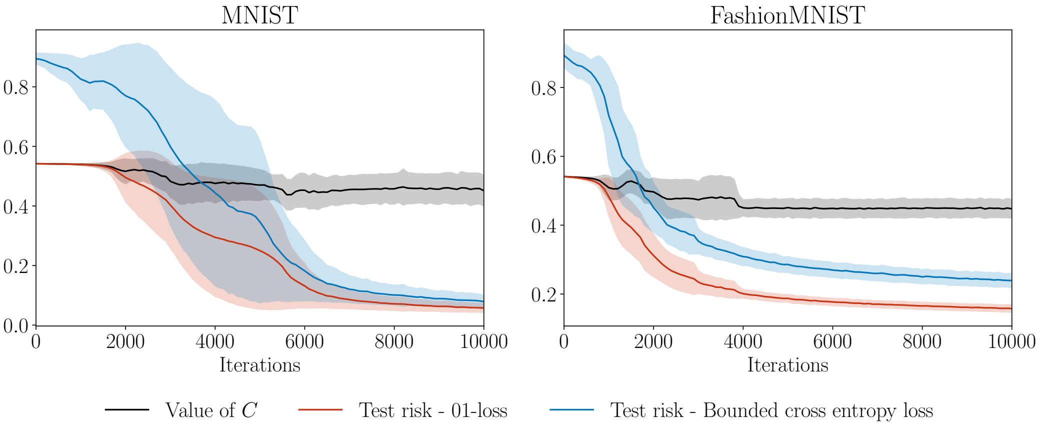
$\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ is *quadratically self-bounded w.r.t.* ℓ and $C > 0$ (namely $\mathtt{QSB}(\ell, C)$) if

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z})^2\right] \leq C\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \; (= C\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z})\right])$$

- QSB intricates $\mathcal{D} \in \mathcal{M}(\mathcal{Z})$ with $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$
- Satisfied if $\ell \in [0, K]$ with $C = K$.
- $\mathtt{QSB}$ quantifies the 'flatness' of the post-training minima reached by Q.

IS THE QSB ASSUMPTION VERIFIED IN PRACTICE?

QSB holds for 3-layer neural nets trained on MNIST (black curve)!



Theorem

For any $C > 0$, data-free prior P, with probability at least $1 - \delta$ for any $m > 0$, and Q being $\texttt{Poinc}(c_P)$, $\texttt{QSB}(\ell, C)$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{Q}) + 2C\frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{m} + \frac{1}{C}c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right].$$

FAST-RATE GENERALISATION BOUNDS VIA FLAT MINIMA (2)

Theorem

For any $C > 0$, data-free prior P, with probability at least $1-\delta$ for any $m > 0$, and Q being $\texttt{Poinc}(c_P)$, $\texttt{QSB}(\ell, C)$,

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathrm{R}}_{\mathcal{S}}(\mathrm{Q}) + 2C\frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{m} + \frac{1}{C}c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right].$$

If $\mathcal{D}$ is also $\texttt{Poinc}$:

With more minor technical assumptions, for any Q being $\texttt{Poinc}(c_P)$ with $\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq C$:

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathrm{R}}_{\mathcal{S}}(\mathrm{Q}) + 2C\frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{m} + \frac{1}{C}\left(c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] + c_P(\mathcal{D}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left(\left\|\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\nabla_z \ell(h,\mathbf{z})]\right\|^2\right)\right).$$

FULLY EMPIRICAL FAST RATE

Current drawback: bounds are not empirical.

FULLY EMPIRICAL FAST RATE

Current drawback: bounds are not empirical.

Solution: $\mathcal{C}^2$ gradient-lipschitz losses!

Theorem

For any $C_1, C_2, c > 0$, with probability at least $1-\delta$, for any $m > 0$, Q being $\texttt{Poinc}(c_P)$ with constant c, $\texttt{QSB}(\ell, C_1)$, $\texttt{QSB}\left(\|\nabla_h \ell\|^2, C_2\right)$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{Q}) + \mathcal{O}\left(\mathbb{E}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\|\nabla_h \ell(h, \mathbf{z}_i)\|^2\right] + \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \log(1/\delta)}{m}\right).$$

TAKE-HOME MESSAGES

If Q satisfies either

1. Flat minima for $\hat{\mathrm{R}}_{\mathcal{S}}$ and $\mathrm{R}_{\mathcal{D}}$,
2. if ℓ gradient-lipschitz, flat minima for $\hat{\mathrm{R}}_{\mathcal{S}}$ and empirical gradient norms,

then Q generalises well!

GIBBS POSTERIORS

Current limitation: with Poincaré posteriors, KL is uncontrolled.

GIBBS POSTERIORS

Current limitation: with Poincaré posteriors, KL is uncontrolled.

Solution: consider Gibbs posterior with log-Sobolev priors!

Definition

$\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}$ is the Gibbs posterior *w.r.t.* prior P with *inverse temperature* $\gamma > 0$ if

$$d\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}(h) \propto \exp\left(-\gamma\hat{\mathsf{R}}_{\mathcal{S}}(h)\right) dP(h)$$

.

Why focus on those?

- Minimise Catoni's bound (Alquier *et al.*, 2016, Theorem 4.1)
- if P `L-Sob`(+ technical assumptions) and $\ell = \ell_1 + \ell_2$ (ℓ_1 convex, twice differentiable, ℓ_2 bounded) then $\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}$ is `L-Sob`.

Theorem

For any $C > 0$, any $\gamma > 0$, any prior P $\texttt{L-Sob}(c_{LS})$ (+ technical assumptions), if $\ell = \ell_1 + \ell_2$ (as above), then with probability at least $1 - \delta$, for any $m > 0$, Q being $\texttt{QSB}(\ell, C)$:

$$\mathsf{R}_{\mathcal{D}}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}) + \mathcal{O}\left(C \frac{\gamma^2\, \mathbb{E}_{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}}\left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}}(h)\|^2\right] + \log(1/\delta)}{m} + \frac{1}{C} \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}} \left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right]\right).$$

Theorem

For any $C > 0$, any $\gamma > 0$, any prior P $\texttt{L-Sob}(c_{LS})$ (+ technical assumptions), if $\ell = \ell_1 + \ell_2$ (as above), then with probability at least $1 - \delta$, for any $m > 0$, Q being $\texttt{QSB}(\ell, C)$:

$$\mathsf{R}_{\mathcal{D}}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}) + \mathcal{O}\left(C \frac{\gamma^2 \, \mathbb{E}_{h\sim \mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}}\left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}}(h)\|^2 \right] + \log(1/\delta)}{m} + \frac{1}{C} \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}} \left[\mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}} \left(\|\nabla_h \ell(h,\mathbf{z})\|^2 \right) \right] \right).$$

KL small if a flat minima on $\hat{\mathsf{R}}_{\mathcal{S}}$ is reached:

$\rightarrow$ **Flat minima fully explain generalisation here!**

1. Gibbs posterior generalises well if they reach a flat minima on both $\hat{R}_{\mathcal{S}}$ and $R_{\mathcal{D}}$.
2. Flatness of the minimum on $\hat{R}_{\mathcal{S}}$ controls the expansion of KL.

Drawback: results hold for probabilistic predictors

Drawback: results hold for probabilistic predictors

Answer: Exploit the 2-Wasserstein distance to obtain guarantees valid for deterministic predictors (Diracs)

Key tool: a novel change of measure inequality

For any f gradient lipschitz, any P, Q:

$$\mathbb{E}_{h\sim \mathrm{Q}}[f(h)] \leq \frac{G}{2} W_2^2(Q,P) + \mathbb{E}_{h\sim \mathrm{P}}[f(h)] + D\mathbb{E}_{h\sim \mathrm{Q}}[\|\nabla f(h)\|].$$

NB: a variant of this formula with a KL is attainable if $\mathrm{Q} << \mathrm{P}$ and P is `L-Sob` !

Assumption

- Gradient-lipschitz loss.
- $\mathrm{P} \propto exp(-V(h))dh$

Theorem

Let $\delta \in (0,1)$ and $P \in \mathcal{M}(\mathcal{H})$ a data-free prior. Assume $\mathcal{H}$ has a finite diameter $D > 0$, $\ell \geq 0$ and that for any m, the generalisation gap $\Delta_{\mathcal{S}_m}$ is G gradient-Lipschitz. Assume that $\mathbb{E}_{h\sim \mathrm{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2] \leq \sigma^2$, then the following holds with probability at least $1-\delta$, for any $m > 0$ and any Q:

$$\mathsf{R}_D(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{G}{2}W_2^2(\mathrm{Q},\mathrm{P}) + \sqrt{\frac{2\sigma^2\log\left(\frac{1}{\delta}\right)}{m}} + D\mathbb{E}_{h\sim\mathrm{Q}}\left(\left\|\nabla_h \mathsf{R}_{\mathcal{D}}(h) - \nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right\|\right)$$

- We mathematically quantify the impact of flat minima in generalisation: momentum in Catoni's bound!
- The `QSB` condition is verified on basic neural nets (classification) with constant C sharper than 1!
- A crucial future lead: understanding why optimisation procedures on deep nets lead to flat minima: **here we are only able to explain why flat minima generalise well, not how we reach them.**

Full paper available at `https://arxiv.org/abs/2402.08508`

REFERENCES I

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*. (2016).

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring Generalization in Deep Learning. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. (2017). URL: `https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html`.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (2020). URL: `https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html`.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. (2020). URL: `https://openreview.net/forum?id=SJgIPJBFvH`.

REFERENCES II

Lucas Liebenwein, Ramin Hasani, Alexander Amini, and Daniela Rus. Sparse flows: Pruning continuous-depth models. *Advances in Neural Information Processing Systems*. 34. (2021).

David Stutz, Matthias Hein, and Bernt Schiele. Relating Adversarially Robust Generalization to Flat Minima. *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. *IEEE*. (2021). DOI: 10.1109/ICCV48922.2021.00771. URL: https://doi.org/10.1109/ICCV48922.2021.00771.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization. *CoRR*. abs/2307.11007. (2023). DOI: 10.48550/ARXIV.2307.11007. arXiv: 2307.11007. URL: https://doi.org/10.48550/arXiv.2307.11007.

Thank you for your attention!