PAC-Bayes-Chernoff bounds for unbounded losses

Ioar Casado, Luis A. Ortega, Aritz Pérez, Andrés R. Masegosa

NeurIPS, December 2024

Introduction

PAC-Bayes theory provides high-probability generalization bounds for randomized learning algorithms.

- A randomized learning algorithms defines probability measure $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ over the set of candidate models $\boldsymbol{\Theta}$.

Notation:

- Data, $D = \{\mathbf{x}_i\}_{i=1}^n$, is i.i.d. generated from an unknown distribution, ν, with support on $\mathcal{X}$
- We have a loss function $\ell : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}_+$
- *Population risk* of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is defined as $L(\boldsymbol{\theta}) := \mathbb{E}_\nu[\ell(\boldsymbol{\theta}, \mathbf{X})]$
- *Empirical risk* of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is defined as $\hat{L}(\boldsymbol{\theta}, D) := \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, \mathbf{x}_i)$

Standard PAC-Bayes bounds for bounded losses (McAllester, 2003):

$$\mathbb{E}_{\rho}[L(\boldsymbol{\theta})] \leq \mathbb{E}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \sqrt{\frac{KL(\rho|\pi) + \log\frac{2\sqrt{n}}{\delta}}{2n}},$$

- The inequality holds simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ with probability no less than $1-\delta$ over the choice of $D \sim \nu^n$.
- We can minimize the bound with respect to ρ to obtain novel learning algorithms.

Introduction

Unbounded losses are widely used in machine learning (e.g., cross-entropy, MSE).

- PAC-Bayes bounds for unbounded losses involve extra difficulties.
- Most existing PAC-Bayes bounds for unbounded losses are derived from next result:

PAC-Bayes (oracle) bound for unbounded losses

For any $\delta \in (0,1)$ and any $\lambda > 0$, with probability at least $1-\delta$ over draws of $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \frac{1}{\lambda}\left[\frac{\mathrm{KL}(\rho|\pi) + \log\frac{f_{\pi,\nu}(\lambda)}{\delta}}{n}\right],$$

where $f_{\pi,\nu}(\lambda) := \mathbb{E}_{\pi}\,\mathbb{E}_{\nu^n}\left[e^{\lambda n\,(L(\boldsymbol{\theta})-\hat{L}(D,\boldsymbol{\theta}))}\right]$.

[Alquier, P., Ridgway, J., & Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. Journal of Machine Learning Research, 17(236), 1-41.]

Main difficulties:

- The exponential moment term $f_{\pi,\nu}(\lambda)$ has to be bounded using extra assumptions on the loss (e.g., sub-Gaussian assumption).
- The free parameter $\lambda > 0$ cannot be exactly optimized (discrete grid + union bounds).

Contributions

- A **novel PAC-Bayes oracle bound** for unbounded losses
  - Extends classic Cramér-Chernoff bounds to the PAC-Bayesian setup.
- Provides a **general framework to obtain empirical bounds** where:
  - **The free parameter λ is exactly optimized** without resorting to union-bound approaches.
  - The exponential moment term is averaged by the posterior, resulting in **more informative generalization bounds**.
  - Can be minimized to obtain **novel posteriors**.
- We illustrate the framework in several cases: generalized sub-Gaussian losses, L2 regularization, and input-gradient regularization.

Definition (Expected Cramér transform)

We define our *comparator* function for any $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ as the *expected Cramér trasform*:

$$\Lambda^{\star}_{\rho}(a) := \sup_{\lambda \in [0,b)} \{\lambda a - \mathop{\mathbb{E}}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]\}, \quad a \in \mathbb{R}\,. \tag{1}$$

Where $\Lambda_{\boldsymbol{\theta}}(\lambda) := \log \mathbb{E}_{\nu}\left[e^{\lambda\,(L(\boldsymbol{\theta})-\ell(\mathbf{x},\boldsymbol{\theta}))}\right]$ is the Cumulant Generating Function (CGF) of the loss.

Main Theorem

Definition (Expected Cramér transform)

We define our *comparator* function for any $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ as the *expected Cramér trasform*:

$$\Lambda^\star_\rho(a) := \sup_{\lambda \in [0,b)} \{\lambda a - \mathbb{E}_\rho[\Lambda_{\boldsymbol{\theta}}(\lambda)]\}, \quad a \in \mathbb{R}\,. \tag{1}$$

Where $\Lambda_{\boldsymbol{\theta}}(\lambda) := \log \mathbb{E}_\nu \left[e^{\lambda\,(L(\boldsymbol{\theta}) - \ell(\mathbf{x},\boldsymbol{\theta}))}\right]$ is the Cumulant Generating Function (CGF) of the loss.

Theorem (PAC-Bayes-Chernoff bound)

For any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of $D \sim \nu^n$,

$$\mathbb{E}_\rho[L(\boldsymbol{\theta})] \leq \mathbb{E}_\rho[\hat{L}(D,\boldsymbol{\theta})] + (\Lambda^\star_\rho)^{-1}\left(\frac{\mathrm{KL}(\rho|\pi) + \log\frac{n}{\delta}}{n-1}\right),$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

Main Theorem

Definition (Expected Cramér transform)

We define our *comparator* function for any $\rho \in \mathcal{M}_1(\mathbf{\Theta})$ as the *expected Cramér trasform*:

$$\Lambda^\star_\rho(a) := \sup_{\lambda \in [0,b)} \{\lambda a - \mathbb{E}_\rho[\Lambda_{\boldsymbol{\theta}}(\lambda)]\}, \quad a \in \mathbb{R}\,. \tag{1}$$

Where $\Lambda_{\boldsymbol{\theta}}(\lambda) := \log \mathbb{E}_\nu \left[e^{\lambda\,(L(\boldsymbol{\theta}) - \ell(\mathbf{x},\boldsymbol{\theta}))}\right]$ is the Cumulant Generating Function (CGF) of the loss.

Theorem (PAC-Bayes-Chernoff bound)

For any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of $D \sim \nu^n$,

$$\mathbb{E}_\rho[L(\boldsymbol{\theta})] \leq \mathbb{E}_\rho[\hat{L}(D,\boldsymbol{\theta})] + (\Lambda^\star_\rho)^{-1}\left(\frac{\mathrm{KL}(\rho|\pi) + \log\frac{n}{\delta}}{n-1}\right),$$

simultaneously for every $\rho \in \mathcal{M}_1(\mathbf{\Theta})$.

Equivalently:

$$\mathbb{E}_\rho[L(\boldsymbol{\theta})] \leq \mathbb{E}_\rho[\hat{L}(D,\boldsymbol{\theta})] + \frac{\mathrm{KL}(\rho|\pi) + \log\frac{n}{\delta}}{\lambda(n-1)} + \frac{\mathbb{E}_\rho[\Lambda_{\boldsymbol{\theta}}(\lambda)]}{\lambda}$$

simultaneously for every $\rho \in \mathcal{M}_1(\mathbf{\Theta})$ and $\lambda \in (0,b)$.

Preliminaries: an auxiliary lemma

Lemma

For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $c \geq 0$, we have

$$\mathbb{P}_{D\sim\nu^n}\Big(n\Lambda^{\star}_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta})) \geq c\Big) \leq \mathbb{P}_{X\sim\exp(1)}\Big(X \geq c\Big).$$

Proof.

Careful rewriting of the Cramér-Chernoff bound after changes of variables. □

Using the fact that for any random variable Z with support $\Omega \subseteq \mathbb{R}_+$ its expectation can be written as

$$\mathbb{E}[Z] = \int_{\Omega} P(Z \geq z)dz\,, \tag{2}$$

the previous lemma will allow us to bound the exponential term $\mathbb{E}_{\nu^n}\left(e^{m\Lambda^{\star}_{\boldsymbol{\theta}}(L(\boldsymbol{\theta})-\hat{L}(D,\boldsymbol{\theta}))}\right)$ in our main theorem.

Main Theorem, proof sketch

Let $m < n$. First, by Jensen,

$$m\Lambda^\star_\rho\left(\mathop{\mathbb{E}}_\rho L(\boldsymbol{\theta}) - \mathop{\mathbb{E}}_\rho \hat{L}(D,\boldsymbol{\theta})\right) \leq m \mathop{\mathbb{E}}_\rho\left[\Lambda^\star_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta}))\right]$$

.

Apply Donsker-Varadhan's lemma on the right side to obtain

$$m\Lambda^\star_\rho\left(\mathop{\mathbb{E}}_\rho L(\boldsymbol{\theta}) - \mathop{\mathbb{E}}_\rho \hat{L}(D,\boldsymbol{\theta})\right) \leq \mathrm{KL}(\rho|\pi) + \log \mathop{\mathbb{E}}_\pi\left(e^{m\Lambda^\star_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta}))}\right)$$

After Markov's inequality + Fubini, with probability at least $1-\delta$,

$$m\Lambda^\star_\rho\left(\mathop{\mathbb{E}}_\rho L(\boldsymbol{\theta}) - \mathop{\mathbb{E}}_\rho \hat{L}(D,\boldsymbol{\theta})\right) \leq \mathrm{KL}(\rho|\pi) + \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_\pi \mathop{\mathbb{E}}_{\nu^n}\left(e^{m\Lambda^\star_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta}))}\right).$$

Using the auxiliary lemma we obtain

$$\mathop{\mathbb{E}}_{\nu^n}\left(e^{m\Lambda^\star_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta}))}\right) \leq \frac{n}{n-m}.$$

Taking $m = n-1$ and applying $\left(\Lambda^\star_\rho\right)^{-1}$ in both sides concludes the proof.

Let us unwrap our bound:

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \inf_{\lambda\in[0,b)} \left\{ \frac{\mathrm{KL}(\rho|\pi) + \log\frac{n}{\delta}}{\lambda(n-1)} + \frac{\mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]}{\lambda} \right\}$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

Observations:

- Parameter-free bound without union-bounds at a $\log n$ cost.
- Generalization of $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ depends on a three-way trade-off.
  - Minimize the empirical Gibbs loss $\mathbb{E}_{\rho}[\hat{L}(D,\boldsymbol{\theta})]$.
  - Minimize the KL term $\mathrm{KL}(\rho|\pi)$.
  - **[Novel term]** Minimize the CGF term $\mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]$
    - Directly related to regularization **(Masegosa&Ortega, 2023)**.

Relation to previous bounds:

- If the loss is the 0-1 loss, we recover Langford-Seeger's bound **(Seeger, 2002)**.

Relation to previous bounds: Bounded CGFs

Bounded CGF assumptions is the standard approach to derived PAC-Bayes bounds for unbounded loss:

- If loss is σ-**sub-gaussian,**

$$\Lambda_{\boldsymbol{\theta}}(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2 \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

- We can recover previous bounds **(Hellstrom & Durisi, 2021)**

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \sqrt{2\sigma^2 \frac{KL(\rho|\pi) + \log\frac{n}{\delta}}{n-1}},$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

Relation to previous bounds: Bounded CGFs

Bounded CGF assumptions is the standard approach to derived PAC-Bayes bounds for unbounded loss:

- If loss is σ-**sub-gaussian,**

$$\Lambda_{\boldsymbol{\theta}}(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2 \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

- We can recover previous bounds **(Hellstrom & Durisi, 2021)**

$$\mathbb{E}_{\rho}[L(\boldsymbol{\theta})] \leq \mathbb{E}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \sqrt{2\sigma^2 \frac{KL(\rho|\pi) + \log\frac{n}{\delta}}{n-1}},$$

  simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

- Uniformly bound $\Lambda_{\boldsymbol{\theta}}(\lambda) \leq \psi(\lambda)$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, necessarily discards information about the statistical properties of individual models.
  - There are models with very different CGF **(Masegosa&Ortega, 2023)**.

Beyond bounded CGFs

Definition (Model-dependent bounded CGF)

A loss function ℓ has model-dependent bounded CGF if for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there is a convex and continuously differentiable function $\psi(\boldsymbol{\theta}, \lambda)$ such that $\psi(\boldsymbol{\theta}, 0) = \psi'(\boldsymbol{\theta}, 0) = 0$ and $\forall \lambda \geq 0$,

$$\Lambda_{\boldsymbol{\theta}}(\lambda) := \log \mathop{\mathbb{E}}_{\nu}\left[e^{\lambda\,(L(\boldsymbol{\theta}) - \ell(\mathbf{x}, \boldsymbol{\theta}))}\right] \leq \psi(\boldsymbol{\theta}, \lambda)\,. \tag{3}$$

Theorem

Let ℓ be a loss function with model-dependent bounded CGF. Let $\pi \in \mathcal{M}_1(\boldsymbol{\Theta})$ be any prior independent of D. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \frac{\mathrm{KL}(\rho|\pi) + \log \frac{n}{\delta}}{\lambda(n-1)} + \frac{\mathbb{E}_\rho[\psi(\boldsymbol{\theta}, \lambda)]}{\lambda}.$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ and $\lambda \in (0, b)$.

Example: sub-Gaussian losses

If the loss function is σ^2-sub-Gaussian, we have $\Lambda_{\boldsymbol{\theta}}(\lambda) \leq \frac{\lambda\sigma^2}{2}$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (**Hellström & Durisi, 2021**):

$$\mathop{\mathbb{E}}_{\rho} L(\boldsymbol{\theta}) \leq \mathop{\mathbb{E}}_{\rho} \hat{L}(D, \boldsymbol{\theta}) + \sqrt{2\sigma^2 \frac{\mathrm{KL}(\rho|\pi) + \log\frac{n}{\delta}}{n-1}}. \tag{4}$$

Our theorem allows the following, more general assumption: $\Lambda_{\boldsymbol{\theta}}(\lambda) \leq \frac{\lambda\sigma(\boldsymbol{\theta})^2}{2}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

$$\mathop{\mathbb{E}}_{\rho} L(\boldsymbol{\theta}) \leq \mathop{\mathbb{E}}_{\rho} \hat{L}(D, \boldsymbol{\theta}) + \sqrt{2 \mathop{\mathbb{E}}_{\rho}[\sigma(\boldsymbol{\theta})^2] \frac{KL(\rho|\pi) + \log\frac{n}{\delta}}{n-1}}, \tag{5}$$

Remark:
The proxy variance σ^2 in equation (4) is a worst-case constant, hence (5) is more general and potentially tighter. It also shows that generalization depends on finding models with smaller proxy variance.

Example: sub-Gaussian losses

If the loss function is σ^2-sub-Gaussian, we have $\Lambda_{\boldsymbol{\theta}}(\lambda) \leq \frac{\lambda\sigma^2}{2}$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (**Hellström & Durisi, 2021**):

$$\mathop{\mathbb{E}}_{\rho} L(\boldsymbol{\theta}) \leq \mathop{\mathbb{E}}_{\rho} \hat{L}(D, \boldsymbol{\theta}) + \sqrt{2\sigma^2 \frac{\mathrm{KL}(\rho|\pi) + \log \frac{n}{\delta}}{n-1}}. \tag{4}$$

Our theorem allows the following, more general assumption: $\Lambda_{\boldsymbol{\theta}}(\lambda) \leq \frac{\lambda\sigma(\boldsymbol{\theta})^2}{2}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

$$\mathop{\mathbb{E}}_{\rho} L(\boldsymbol{\theta}) \leq \mathop{\mathbb{E}}_{\rho} \hat{L}(D, \boldsymbol{\theta}) + \sqrt{2 \mathop{\mathbb{E}}_{\rho}[\sigma(\boldsymbol{\theta})^2] \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1}}, \tag{5}$$

Remark:
The proxy variance σ^2 in equation (4) is a worst-case constant, hence (5) is more general and potentially tighter. It also shows that generalization depends on finding models with smaller proxy variance.

However, **we are not limited to using tail assumptions!**

L2 regularization

L2 regularization minimizes an objective function of the form

$$\hat{L}(D, \boldsymbol{\theta}) + k\|\boldsymbol{\theta}\|_2^2,$$

where $k > 0$ is a trade-off parameter.

If the loss $\ell(\mathbf{x}, \boldsymbol{\theta})$ is M-Lipschitz with respect to $\boldsymbol{\theta}$, as shown in **(Masegosa&Ortega, 2023)**,

$$\Lambda_{\boldsymbol{\theta}}(\lambda) \leq 2M\lambda^2\|\boldsymbol{\theta}\|_2^2. \tag{6}$$

Theorem

If ℓ satisfies the inequality above, then for any $\delta_1 \in (0, 1)$, with probability at least $1 - \delta_1$ over draws $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \sqrt{2M \mathop{\mathbb{E}}_{\rho}\left[\|\boldsymbol{\theta}\|_2^2\right] \frac{\mathrm{KL}(\rho|\pi) + \log \frac{n}{\delta_1}}{n-1}} \tag{7}$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

Input-gradient regularization through PAC-Bayes

Input-gradient regularization (**Varga et al., 2017**) minimizes an objective function of the form

$$\hat{L}(D, \boldsymbol{\theta}) + k \frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\mathbf{x}} \ell(\mathbf{x}_i, \boldsymbol{\theta})\|_2^2,$$

where $k > 0$ is a trade-off parameter. This approach is often used to make models more robust against disturbances in input data and adversarial attacks (**Ross & Doshi-Velez, 2018**). With the proper assumptions, our bounds provide a **PAC-Bayesian interpretation**.

The connection is provided by the assumption that the underlying distribution satisfies a log-Sobolev inequality (**Chafaï, 2004**):

$$\Lambda_{\boldsymbol{\theta}}(\lambda) \leq \frac{M}{2} \lambda^2 \mathop{\mathbb{E}}_{\nu} \|\nabla_{\mathbf{x}} \ell(\mathbf{x}, \boldsymbol{\theta})\|_2^2 \tag{8}$$

for every $\lambda > 0$ and some $M > 0$.

Input-gradient regularization through PAC-Bayes

Theorem (Oracle PAC-Bayes bound for input-gradients)

If ν satisfies the inequality above, then for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$ over draws $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \sqrt{2M \mathop{\mathbb{E}}_{\rho}\left[\mathop{\mathbb{E}}_{\nu} \|\nabla_{\mathbf{x}} \ell(\mathbf{x},\boldsymbol{\theta})\|_2^2\right] \frac{\mathrm{KL}(\rho|\pi) + \log \frac{n}{\delta_1}}{n-1}} \tag{9}$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

Input-gradient regularization through PAC-Bayes

Theorem (Oracle PAC-Bayes bound for input-gradients)

If ν satisfies the inequality above, then for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$ over draws $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \sqrt{2M \mathop{\mathbb{E}}_{\rho}\left[\mathop{\mathbb{E}}_{\nu} \|\nabla_{\mathbf{x}} \ell(\mathbf{x},\boldsymbol{\theta})\|_2^2\right] \frac{\mathrm{KL}(\rho|\pi) + \log \frac{n}{\delta_1}}{n-1}} \tag{9}$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

We can obtain an empirical bound if we assume that the gradient norms have sub-Gaussian tails.

Theorem (Empirical PAC-Bayes bound for input-gradients)

With the same conditions as above, assume $\|\nabla_{\mathbf{x}} \ell(\mathbf{x},\boldsymbol{\theta})\|_2^2$ is σ^2-sub-Gaussian for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \sqrt{2M \left(\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\rho}\left[\|\nabla_{\mathbf{x}} \ell(\mathbf{x}_i,\boldsymbol{\theta})\|_2^2\right] + \frac{\sigma^2}{2}\right) \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1}}$$
$$+ \sqrt{2M} \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1}.$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

If we fix $\lambda > 0$ and repeat the same procedure —log-Sobolev + sub-Gaussianity of gradients— starting with the parametric bound

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \frac{\mathrm{KL}(\rho|\pi) + \log\frac{n}{\delta}}{\lambda(n-1)} + \frac{\mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]}{\lambda},$$

the subsequent bound can be minimized w.r.t. $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$, resulting in the optimal posterior:

$$\rho^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\Big\{ -(n-1)\big(\hat{L}(D,\boldsymbol{\theta}) + k\frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\mathbf{x}}\ell(\mathbf{x}_i,\boldsymbol{\theta})\|_2^2\big)\Big\},$$

where $k = \lambda\frac{M}{2}$. The optimal posterior concentrates its mass of probability in models minimizing the term $\hat{L}(D,\boldsymbol{\theta}) + \frac{\lambda M}{2}\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}}\ell(\mathbf{x}_i,\boldsymbol{\theta})\|_2^2$, which is exactly the minimization objective of input-gradient regularization with trade-off parameter $\frac{\lambda M}{2}$.

Conclusions

- **Novel PAC-Bayes Oracle Bound**: Introduced a novel PAC-Bayes oracle bound leveraging the Cramer transform.
- **Optimization of Free Parameter** $(\lambda > 0)$: Facilitates exact optimization of the free parameter $\lambda > 0$, solving a longstanding issue in PAC-Bayesian methods and allowing for tighter empirical bounds.
- **Introduction of Model-Dependent Assumptions**: Enables flexible, richer model-dependent assumptions for bounding the Cumulative Generating Function (CGF).
    - **Applicability Across Diverse Assumptions**: Demonstrates the bound's utility through assumptions such as generalized sub-Gaussian losses, parameter norm bounds, and log-Sobolev inequalities, including a novel empirical PAC-Bayes bound.

Thanks for your attention!! :)