Information, Concentration, and Learning

Maxim Raginsky

University of Illinois at Urbana-Champaign

North American Summer School on Information Theory
Boston University, July 2019

Part I

Background, Definitions, Key Results

Learning: Data-Driven Stochastic Optimization

- Goal: stochastic optimization

$$\text{minimize} \qquad L_\mu(w) := \mathbf{E}_\mu[\ell(w,Z)] = \int_{\mathsf{Z}} \ell(w,z)\mu(\mathrm{d}z)$$

  where:

  - w is an element of the *hypothesis space* W
  - Z is a random element of the *instance space* Z
  - $\mu := \mathcal{L}(Z)$ is unknown
  - $\ell : \mathsf{W} \times \mathsf{Z} \to \mathbb{R}_+$ is the *loss function*

- $L_\mu(w)$ is the *population loss* of the hypothesis w w.r.t. μ
- Data-driven optimization: μ is unknown, but we have access to *training data*

$$\boldsymbol{Z} = (Z_1, \ldots, Z_n), \qquad Z_i \overset{\text{i.i.d.}}{\sim} \mu$$

Learning Algorithms and their Performance

- Given: training data $\boldsymbol{Z} \sim \mu^{\otimes n}$
- Learning algorithm: a stochastic transformation (channel) from training data to hypotheses:

$$\boldsymbol{Z} \xrightarrow{P_{W|\boldsymbol{Z}}} W$$

  where W is a *random* element of the hypothesis space W
- Goal of learning (broadly speaking): design $P_{W|\boldsymbol{Z}}$, such that the *out-of-sample loss*

$$L_\mu(W) = \int_{\mathsf{Z}} \ell(W, z)\mu(\mathrm{d}z)$$

  is suitably small (either in expectation or with high probability)
- ***Caution!!*** $L_\mu(W)$ is a *random variable*

Examples: I

1. Binary classification
   - $\mathsf{Z} = \mathsf{X} \times \{0,1\}$
   - each $w \in \mathsf{W}$ corresponds to a *classifier* $f_w : \mathsf{X} \to \{0,1\}$
   - $\ell(w,z) = \ell(w,(x,y)) = \mathbf{1}_{\{y \neq f_w(x)\}}$
   - $L_\mu(w) = \mathbf{P}_\mu[Y \neq f_w(X)]$
2. Regression with quadratic loss
   - $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$, where $\mathsf{Y} \subseteq \mathbb{R}$
   - each $w \in \mathsf{W}$ corresponds to a *predictor* $f_w : \mathsf{X} \to \mathbb{R}$
   - $\ell(w,z) = \ell(w,(x,y)) = (y - f_w(x))^2$
   - $L_\mu(w) = \mathbf{E}_\mu[(Y - f_w(X))^2]$

These are examples of *supervised learning problems*: the instance z splits into a ‘feature’ x and a ‘label’ y, and the goal of learning is to predict Y given X when $(X,Y) \sim \mu$.

Examples: II

3. Clustering
   - Z is a metric space with metric ρ
   - $\mathsf{W} = \mathsf{Z}^k$
   - $\ell(w, z) = \ell((w_1, \dots, w_k), z) = \min_{1 \le j \le k} \rho(w_j, z)^p$, $p \ge 1$
   - $L_\mu(w) = \mathbf{E}_\mu \left[\min_{1 \le j \le k} \rho(w_j, Z)^p\right]$
4. Density estimation
   - $\mathsf{Z} \subseteq \mathbb{R}^d$
   - each $w \in \mathsf{W}$ corresponds to a *probability density* f_w on $\mathbb{R}^d$
   - $\ell(w, z) = -\log f_w(z)$
   - $L_\mu(w) = \mathbf{E}_\mu[-\log f_w(Z)]$

These are examples of *unsupervised learning problems*: the goal is to find some ‘structure’ in the probability space (Z, μ).

Empirical Loss and Generalization Error

- The data-generating distribution μ is unknown; how do we evaluate the quality of the learned hypothesis W?
- Empirical loss of a fixed hypothesis $w \in \mathsf{W}$:

$$L_{\boldsymbol{Z}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$$

— unbiased estimate of $L_\mu(w)$, $\mathbf{E}[L_{\boldsymbol{Z}}(w)] = L_\mu(w)$

- Empirical loss of W (a.k.a. *resubstitution estimate*)

$$L_{\boldsymbol{Z}}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i)$$

can be computed from the available information $(\boldsymbol{Z}, W)$, but is a biased estimate: $\mathbf{E}[L_{\boldsymbol{Z}}(W)] \neq \mathbf{E}[L_\mu(W)]$

Generalization error:

$$\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) := \mathbf{E}[L_\mu(W) - L_{\boldsymbol{Z}}(W)]$$

What Does $\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})$ Tell Us?

Suppose there exists an *optimal hypothesis* $w_{\mathsf{opt}} \in \mathsf{W}$:

$$L_\mu(w_{\mathsf{opt}}) = \min_{w \in \mathsf{W}} L_\mu(w)$$

Let us analyze the expected *excess risk* of $P_{W|\boldsymbol{Z}}$ w.r.t. μ:

$$\begin{aligned}
\mathsf{ex}(\mu, P_{W|\boldsymbol{Z}}) &:= \mathbf{E}[L_\mu(W)] - L_\mu(w_{\mathsf{opt}}) \\
&= \mathbf{E}[L_\mu(W) - L_{\boldsymbol{Z}}(W)] + \mathbf{E}[L_{\boldsymbol{Z}}(W)] - L_\mu(w_{\mathsf{opt}}) \\
&= \mathbf{E}[L_\mu(W) - L_{\boldsymbol{Z}}(W)] + \mathbf{E}[L_{\boldsymbol{Z}}(W) - L_{\boldsymbol{Z}}(w_{\mathsf{opt}})] \\
&= \mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) + \mathbf{E}[L_{\boldsymbol{Z}}(W) - L_{\boldsymbol{Z}}(w_{\mathsf{opt}})]
\end{aligned}$$

Thus, $\mathsf{ex}(\mu, P_{W|\boldsymbol{Z}})$ will be small if:

- $\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})$ is small (i.e., learning algo generalizes well *on average*)
- the empirical risks of W and w_{opt} are close on average

Empirical Risk Minimization and Uniform Convergence

- Empirical Risk Minimization (ERM): $W \in \arg\min_{w \in \mathsf{W}} L_{\boldsymbol{Z}}(w)$
- This leads to the following upper bound:

$$\begin{aligned}
\mathsf{ex}(\mu, P^{\mathsf{ERM}}_{W|\boldsymbol{Z}}) &= \mathsf{gen}(\mu, P^{\mathsf{ERM}}_{W|\boldsymbol{Z}}) + \mathbf{E}\Big[\underbrace{\min_{w \in \mathsf{W}} L_{\boldsymbol{Z}}(w) - L_{\boldsymbol{Z}}(w_{\mathsf{opt}})}_{\le 0}\Big] \\
&\le \mathsf{gen}(\mu, P^{\mathsf{ERM}}_{W|\boldsymbol{Z}}) \\
&\le \mathbf{E}\Big[\sup_{w \in \mathsf{W}} \big|L_{\boldsymbol{Z}}(w) - L_\mu(w)\big|\Big]
\end{aligned}$$

- 'Classical' stat. learning theory: If the *induced function class* $\mathcal{F}_{\ell,\mathsf{W}} := \{\ell(w,\cdot) : w \in \mathsf{W}\}$ is not 'too rich,' then

$$\mathbf{E}\Big[\sup_{w \in \mathsf{W}} \big|L_{\boldsymbol{Z}}(w) - L_\mu(w)\big|\Big] \le \frac{C}{\sqrt{n}},$$

where C measures the complexity of $\mathcal{F}_{\ell,\mathsf{W}}$ and does not depend on μ (distribution-free bound)

Uniform Convergence and Generalization

We can *always* bound the generalization error as

$$\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) \le \mathbf{E}\Big[\sup_{w \in \mathsf{W}} \big|L_{\boldsymbol{Z}}(w) - L_\mu(w)\big|\Big]$$

... but this bound:

- relies on restricting the complexity of the hypothesis space;
- ignores the details of the interaction between the data $\boldsymbol{Z}$ and the algo. output W;
- in particular, may be too conservative if the algo. does not explore the entire W due to fixed computational budget.

Learning does not require uniform convergence: One can construct examples of (ℓ, W), where uniform convergence does not hold (the upper bound does not converge to 0 as $n \to \infty$), yet learning still takes place (Shalev-Shwartz et al., 2010)

So, What Does $\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})$ Really Tell Us?

Let's try one more time:

- consider generating *two* independent training samples

$$\boldsymbol{Z} \sim \mu^{\otimes n}, \qquad \boldsymbol{Z}' \sim \mu^{\otimes m}, \qquad \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Z}'$$

- for each $i \in [n]$, replace Z_i with Z_i':

$$\begin{tikzcd}(Z_1, \dots, Z_{i-1}, {\color{blue}Z_i}, Z_{i+1}, \dots, Z_n) \arrow[r, "P_{W|\boldsymbol{Z}}"] \arrow[d] & W \\ (Z_1, \dots, Z_{i-1}, {\color{red}Z_i'}, Z_{i+1}, \dots, Z_n) \arrow[r, "P_{W|\boldsymbol{Z}}"] & W^{(i)}\end{tikzcd}$$

- Since $(W, Z_1, \dots, Z_i, \dots, Z_n, Z_i') \overset{\mathrm{d}}{=} (W^{(i)}, Z_1, \dots, Z_i', \dots, Z_n, Z_i)$,

$$\begin{aligned}\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[\ell(W, Z_i') - \ell(W, Z_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[\ell(W, Z_i') - \ell(W^{(i)}, Z_i')]\end{aligned}$$

— quantifies the sensitivity of $P_{W|\boldsymbol{Z}}$ to *local modifications* of $\boldsymbol{Z}$

Generalization as Algorithmic Stability

For *any* learning algo. $P_{W|\boldsymbol{Z}}$ and *any* data distribution μ,

$$\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[\ell(W, Z_i') - \ell(W^{(i)}, Z_i')]$$

measures *stability* of $P_{W|\boldsymbol{Z}}$ w.r.t. the training sample.

Rich history, many definitions:

- first introduced by Rogers and Wagner (1978), Devroye and Wagner (1979)
- the term 'algorithmic stability' first used by Kearns and Ron (1999)
- stability goes mainstream: Bousquet and Elisseeff (2002); Poggio et al. (2004); Rakhlin et al. (2005); Shalev-Shwartz et al. (2010)
- mushrooming: Kutin and Niyogi (2002) propose *twelve* (!) different notions of stability

An Example: Uniform Stability

For *any* learning algo. $P_{W|\boldsymbol{Z}}$ and *any* data distribution μ,

$$\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[\ell(W, Z'_i) - \ell(W^{(i)}, Z'_i)]$$

measures *stability* of $P_{W|\boldsymbol{Z}}$ w.r.t. the training sample.

Definition (Bousquet and Elisseeff): $P_{W|\boldsymbol{Z}}$ is ε*-uniformly stable* if

$$\forall i \in [n] : \quad \sup_{z \in \mathsf{Z}} \mathbf{E}[\ell(W, z) - \ell(W^{(i)}, z) | \boldsymbol{Z}, \boldsymbol{Z}'] \le \varepsilon.$$

If $P_{W|\boldsymbol{Z}}$ is ε-uniformly stable, then

$$\begin{aligned}\mathbf{E}[\ell(W, Z'_i)] - \mathbf{E}[\ell(W^{(i)}, Z'_i)] &= \mathbf{E}[\mathbf{E}[\ell(W, Z'_i) - \ell(W^{(i)}, Z'_i) | \boldsymbol{Z}, \boldsymbol{Z}']] \\ &\le \mathbf{E}\left[\sup_{z \in \mathsf{Z}} \mathbf{E}[\ell(W, z) - \ell(W^{(i)}, z) | \boldsymbol{Z}, \boldsymbol{Z}']\right] \le \varepsilon \\ &\Longrightarrow \mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) \le \varepsilon\end{aligned}$$

Distributional Stability

- Uniform stability:

$$\forall i \in [n] : \quad \sup_{z \in \mathsf{Z}} \mathbf{E}[\ell(W, z) - \ell(W^{(i)}, z) | \boldsymbol{Z}, \boldsymbol{Z}'] \le \varepsilon.$$

  — really, a statement about the *conditional distributions* $P_{W|\boldsymbol{Z}=\boldsymbol{z}}$ and $P_{W|\boldsymbol{Z}=\boldsymbol{z}'}$ when $d_{\mathsf{H}}(\boldsymbol{z}, \boldsymbol{z}') = 1$, where

$$d_{\mathsf{H}}(\boldsymbol{z}, \boldsymbol{z}') := \sum_{i=1}^{n} \mathbf{1}_{\{z_i \neq z_i'\}}$$

  is the *Hamming distance* between $\boldsymbol{z}$ and $\boldsymbol{z}'$

- This is very reminiscent of *differential privacy*:

$$d_{\mathrm{H}}(\boldsymbol{z}, \boldsymbol{z}') = 1 \quad \Longrightarrow \quad \mathsf{dist}(P_{W|\boldsymbol{Z}=\boldsymbol{z}}, P_{W|\boldsymbol{Z}=\boldsymbol{z}'}) \le \varepsilon$$

  for a suitably chosen $\mathsf{dist}(\cdot, \cdot)$.

- We will come back to this later.

Information-Theoretic Stability

- Key idea: a learning algo. $P_{W|\boldsymbol{Z}}$ is stable if its output W does not depend 'too much' on any particular element of $\boldsymbol{Z}$.
- Let's make this precise:
  - how do we capture dependence?
  - what does 'not too much' mean?
- Two information-theoretic measures of dependence:

$$\text{mutual information: } I(\boldsymbol{Z};W) = \sum_{i=1}^{n} I(W;Z_i|Z^{i-1})$$

$$\text{erasure information: } I^{-}(\boldsymbol{Z};W) := \sum_{i=1}^{n} I(W;Z_i|\boldsymbol{Z}^{-i})$$

$$\boldsymbol{Z}^{-i} := (Z_1,\dots,Z_{i-1},Z_{i+1},\dots,Z_n)$$

(Verdú and Weissman, 2008)

- To show: if $I(\boldsymbol{Z};W)$ or $I^{-}(\boldsymbol{Z};W)$ is small, then so is $\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})$.

Mutual Information vs. Erasure Information

- Note the conditioning:

$$\text{mutual information: } I(\boldsymbol{Z};W)=\sum_{i=1}^{n} I(W;Z_i|Z^{i-1})$$

$$\text{erasure information: } I^{-}(\boldsymbol{Z};W)=\sum_{i=1}^{n} I(W;Z_i|\boldsymbol{Z}^{-i})$$

- In general, I can be larger or smaller than I^{-}

Lemma. $I(\boldsymbol{Z};W)\le I^{-}(\boldsymbol{Z};W)$

Proof: for any $i\in[n]$

$$\begin{aligned} I(W;Z_i|Z^{i-1}) &= I(W,Z^{i-1};Z_i) && \text{(chain rule, independence of } Z_i\text{'s)} \\ &\le I(W;\boldsymbol{Z}^{-i};Z_i) && \text{(data processing)} \\ &= I(W;Z_i|\boldsymbol{Z}^{-i}) && \text{(chain rule, independence of } Z_i\text{'s)} \end{aligned}$$

then sum over i.

Information-Theoretic Stability: Definitions

Definition. A learning algorithm $P_{W|\boldsymbol{Z}}$ is ε-*stable* w.r.t. data-generating distibution μ in mutual information (resp., erasure information) if

$$I(\boldsymbol{Z};W) \leq n\varepsilon \qquad \text{(resp., } I^{-}(\boldsymbol{Z};W) \leq n\varepsilon\text{)}.$$

- Since $I(\boldsymbol{Z};W) \leq I^{-}(\boldsymbol{Z};W)$, stability in erasure information implies stability in mutual information.
- On the other hand, it is often easier to establish stability in erasure information (e.g., via notions related to differential privacy).

Sufficient Condition for Stability in Erasure Information

Definition (Bassily et al.) Learning algo. $P_{W|\boldsymbol{Z}}$ is ε-*KL-stable* if

$$d_{\mathsf{H}}(\boldsymbol{z}, \boldsymbol{z}') = 1 \qquad \Longrightarrow \qquad D(P_{W|\boldsymbol{Z}=\boldsymbol{z}} \| P_{W|\boldsymbol{Z}=\boldsymbol{z}'}) \leq \varepsilon.$$

Remarks:

1. The definition does not involve μ, only the algo.
2. In the original paper of Bassily et al. (2015), ε is replaced by $2\varepsilon^2$ (to relate to usual notions of differential privacy).

Lemma. If $P_{W|\boldsymbol{Z}}$ is ε-KL-stable, then it is ε-stable in erasure information (and therefore in mutual information) for any μ.

Sufficient Condition for Stability in Erasure Information

Proof.

1. For any $i \in [n]$,

$$I(W; Z_i | \boldsymbol{Z}^{-i}) = \int_{\mathsf{Z}^n} \mu^{\otimes n}(\mathrm{d}\boldsymbol{z}) D(P_{W|\boldsymbol{Z}=\boldsymbol{z}} \| P_{W|\boldsymbol{Z}^{-i}=\boldsymbol{z}^{-i}})$$

2. Marginalization:

$$\begin{aligned} P_{W|\boldsymbol{Z}^{-i}=\boldsymbol{z}^{-i}}(\cdot) &= \int_{\mathsf{Z}} \mu(\mathrm{d}z_i) P_{W|\boldsymbol{Z}=(z_1,\ldots,z_i,\ldots,z_n)}(\cdot) \\ &= \int_{\mathsf{Z}} \mu(\mathrm{d}z') P_{W|\boldsymbol{Z}=\boldsymbol{z}^{i,z'}}(\cdot) \end{aligned}$$

   where $\boldsymbol{z}^{i,z'} := (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_n)$

3. By convexity of the relative entropy:

$$D(P_{W|\boldsymbol{Z}=\boldsymbol{z}} \| P_{W|\boldsymbol{Z}^{-i}=\boldsymbol{z}^{-i}}) \le \int_{\mathsf{Z}} \mu(\mathrm{d}z') D(P_{W|\boldsymbol{Z}=\boldsymbol{z}} \| P_{W|\boldsymbol{Z}=\boldsymbol{z}^{i,z'}}) \le \varepsilon$$

4. $I^{-}(\boldsymbol{Z}; W) = \sum_{i=1}^{n} I(W; Z_i | \boldsymbol{Z}^{-i}) \le n\varepsilon.$

Moment-Generating Functions

Definition. The *logarithmic moment-generating function* of a random variable U is

$$\psi(\lambda) := \log \mathbf{E}[e^{\lambda(U-\mathbf{E}U)}], \qquad \lambda \in \mathbb{R}.$$

Properties:

1. $\psi(\lambda)$ is C^∞ and convex in λ
2. $\psi(0) = \psi'(0) = 0$
3. Exponential Markov inequality: for any $t > 0$,

   $$\mathbf{P}[U \geq \mathbf{E}U + t] = \mathbf{P}[e^{\lambda(U-\mathbf{E}U)} \geq e^{\lambda t}] \leq e^{-(\lambda t - \psi(\lambda))}$$

4. Chernoff bound:

   $$\mathbf{P}[U \geq \mathbf{E}U + t] \leq e^{-\psi^*(t)},$$

   where $\psi^*(t) := \sup_{\lambda \geq 0}\{\lambda t - \psi(\lambda)\}$ is the *Legendre dual* of ψ

Subgaussian Random Variables

- If $U \sim \mathcal{N}(0, \sigma^2)$, then

$$\log \mathbf{E}[e^{\lambda U}] = \frac{\lambda^2 \sigma^2}{2}$$

- We say that a random variable U with $\mathbf{E}U < \infty$ is σ^2-*subgaussian* if

$$\log \mathbf{E}[e^{\lambda(U - \mathbf{E}U)}] \leq \frac{\lambda^2 \sigma^2}{2}, \qquad \lambda \in \mathbb{R}.$$

- Classic example (Hoeffding's lemma): if $-\infty < a \leq U \leq b < \infty$ almost surely, then

$$\log \mathbf{E}[e^{\lambda(U - \mathbf{E}U)}] \leq \frac{\lambda^2 (b-a)^2}{8}, \qquad \forall \lambda \in \mathbb{R}$$

  Thus, any such U is $\frac{(b-a)^2}{4}$-subgaussian.

A Decoupling Estimate

Proposition. Let U and V be two jointly distributed random objects, and let $f(U,V)$ be a real-valued function such that

$$\begin{aligned} \sup_u \log \mathbf{E}[e^{\lambda(f(u,V)-\mathbf{E}[f(u,V)])}] &\le \psi_+(\lambda), \qquad \lambda > 0 \\ \sup_u \log \mathbf{E}[e^{\lambda(f(u,V)-\mathbf{E}[f(u,V)])}] &\le \psi_-(-\lambda), \qquad \lambda < 0 \end{aligned}$$

where ψ_+, ψ_- are convex and $\psi_\pm(0) = \psi'_\pm(0) = 0$. Then

$$\begin{aligned} \mathbf{E}[f(U,V) - f(\bar{U},\bar{V})] &\le \psi_+^{*-1}(I(U;V)), \\ \mathbf{E}[f(\bar{U},\bar{V})] - f(U,V)] &\le \psi_-^{*-1}(I(U;V)) \end{aligned}$$

where:

- $P_{\bar{U},\bar{V}} = P_U \otimes P_V$
- $\psi_\pm^{*-1}$ is the inverse of the Legendre dual $\psi_\pm^*$

Proof

1. Donsker-Varadhan duality: for any $\lambda > 0$,

$$\begin{aligned} D(P_{V|U=u} \| P_V) &\ge \lambda \mathbf{E}[f(u,V)|U=u] - \log \mathbf{E}[e^{\lambda f(u,V)}] \\ &\ge \lambda \big(\mathbf{E}[f(u,V)|U=u] - \mathbf{E}[f(u,V)] \big) - \psi_+(\lambda) \end{aligned}$$

2. Rearrange and optimize:

$$\begin{aligned} \mathbf{E}[f(u,V)|U=u] - \mathbf{E}[f(u,V)] &\le \inf_{\lambda > 0} \frac{D(P_{V|U=u} \| P_V) + \psi_+(\lambda)}{\lambda} \\ &= \psi_+^{*-1}(D(P_{V|U=u} \| P_V)) \end{aligned}$$

   (see, e.g., the book of Boucheron–Lugosi–Massart)

3. Average w.r.t. $U \sim P_U$:

$$\begin{aligned} \mathbf{E}[f(U,V)] - \mathbf{E}[f(\bar{U},\bar{V})] &\le \int P_U(\mathrm{d}u)[\psi_+^{*-1}(D(P_{V|U=u}) \| P_V)] \\ &\le \psi_+^{*-1}(I(U;V)), \end{aligned}$$

   where we have used the fact that ψ_+^{*-1} is concave (since ψ_+^* is convex).

4. The case with $\lambda < 0$ is analogous.

Bounding $\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})$ via Mutual Information

Theorem. Suppose that there exist convex functions $\psi_\pm : \mathbb{R}_+ \to \mathbb{R}$ satisfying $\psi_\pm(0) = \psi'_\pm(0) = 0$, such that

$$\sup_{w \in \mathsf{W}} \mathbf{E}[e^{\pm\lambda(\ell(w,Z)) - \mathbf{E}[\ell(w,Z)]}] \le \psi_\pm(\pm\lambda), \qquad \lambda > 0.$$

Then, for any learning algorithm $P_{W|\boldsymbol{Z}}$ such that $I(W : \boldsymbol{Z}) < \infty$,

$$\psi_+^{*-1}\left(\frac{1}{n} I(W; \boldsymbol{Z})\right) \le \mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) \le \psi_-^{*-1}\left(\frac{1}{n} I(W; \boldsymbol{Z})\right).$$

Remarks:

1. The subgaussian case is due to Xu–Raginsky (2017); related results by Russo–Zou (2016).
2. The general case was analyzed by Jiao–Han–Weissman (2018); Bu-Zou-Veeravalli (2019).

Proof

1. Since $Z_i \overset{\text{i.i.d.}}{\sim} \mu$, for any $w \in \mathsf{W}$ and any $\lambda > 0$

$$\begin{aligned}
&\log \mathbf{E}\left[\exp\left\{\pm\lambda\left(L_{\boldsymbol{Z}}(w) - L_\mu(w)\right)\right\}\right] \\
&= n \log \mathbf{E}\left[\exp\left\{\frac{\pm\lambda}{n}\left(\ell(w,Z) - \mathbf{E}[\ell(w,Z)]\right)\right\}\right] \le n\psi_\pm(\pm\lambda/n)
\end{aligned}$$

2. Now take $U = W$, $V = \boldsymbol{Z}$, $\ell(U,V) = L_{\boldsymbol{Z}}(W)$:

$$\mathbf{E}[f(U,V)] = \mathbf{E}[L_{\boldsymbol{Z}}(W)], \qquad \mathbf{E}[f(\bar{U},\bar{V})] = \mathbf{E}[L_\mu(W)].$$

Apply the Decoupling Estimate to get

$$\begin{aligned}
\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) &\le \inf_{\lambda>0} \frac{I(W;\boldsymbol{Z}) + n\psi_-(\lambda/n)}{\lambda} \\
&= \inf_{\lambda>0} \frac{\frac{1}{n}I(W;\boldsymbol{Z}) + \psi_-(\lambda)}{\lambda} \\
&= \psi_-^{*-1}\left(\frac{1}{n}I(W;\boldsymbol{Z})\right)
\end{aligned}$$

3. The lower bound is similar.

Subgaussian Case

- When $\ell(w, Z)$ is σ^2-subgaussian for every $w \in \mathsf{W}$, we can take

$$\psi_{\pm}(t) = \frac{t^2\sigma^2}{2}, \qquad \forall t \in \mathbb{R}$$

$$\psi_{\pm}^{*-1}(r) = \inf_{\lambda > 0} \frac{r + \lambda^2\sigma^2/2}{\lambda} = \sqrt{2r\sigma^2}$$

- Under the above assumption, for any learning algo. $P_{W|\boldsymbol{Z}}$ we have

$$|\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})| \le \sqrt{\frac{2\sigma^2}{n} I(W; \boldsymbol{Z})}$$

- In particular, if $P_{W|\boldsymbol{Z}}$ is ε-stable w.r.t. μ in mutual information or in erasure information,

$$|\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})| \le \sqrt{2\sigma^2\varepsilon}.$$

A Concentration Inequality for $|L_{\boldsymbol{Z}}(W) - L_\mu(W)|$

- So far, we have been concerned with

$$\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) = \mathbf{E}[L_\mu(W) - L_{\boldsymbol{Z}}(W)]$$

- What about $\mathbf{P}[|L_\mu(W) - L_{\boldsymbol{Z}}(W)| > \varepsilon]$?
- Let's consider an extreme (and boring) case: $W \perp\!\!\!\perp \boldsymbol{Z}$ — the learning algo. just ignores the data.

  Then, assuming $\ell(w, Z)$ is σ^2-subgaussian for all w,

$$\mathbf{P}\left[|L_{\boldsymbol{Z}}(W) - L_\mu(W)| > \varepsilon\right] \le 2e^{-\frac{n\varepsilon^2}{2\sigma^2}}, \qquad \forall \varepsilon > 0$$

  — that is, given $\varepsilon > 0$ and $0 < \delta \le 1$, a sample size $n = \Omega(\frac{2\sigma^2}{\varepsilon^2} \log \frac{2}{\delta})$ suffices to guarantee

$$|L_{\boldsymbol{Z}}(W) - L_\mu(W)| \le \varepsilon \qquad \text{with prob. at least } 1 - \delta.$$

- What happens if $I(W; \boldsymbol{Z})$ is suitably 'small?'

A Concentration Inequality for $|L_{\boldsymbol{Z}}(W) - L_\mu(W)|$

Theorem (Xu–Raginsky). Suppose $\ell(w, Z)$ is σ^2-subgaussian for all $w \in \mathsf{W}$. Let $P_{W|\boldsymbol{Z}}$ be a learning algo. with $I(W; \boldsymbol{Z}) < \infty$. Let $\varepsilon > 0$ and $0 < \delta \le 1$ be given . Then, provided

$$n \ge \frac{8\sigma^2}{\varepsilon^2}\left(\frac{I(W;\boldsymbol{Z})}{\delta} + \log\frac{2}{\delta}\right),$$

we will have

$$\mathbf{P}\left[|L_{\boldsymbol{Z}}(W) - L_\mu(W)| > \varepsilon\right] \le \delta.$$

Remarks:

1. The proof uses the *monitor technique* of Bassily et al.: run the algo. on m independent datasets $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_m$, then select the output with the largest value of $|L_\mu(W_j) - L_{\boldsymbol{Z}_j}(W_j)|$; the resulting ‘super-algo.’ has bounded mutual information.
2. The theorem does not give a ‘high-probability’ bound, due to $\frac{1}{\delta}$ additive term. Bassily et al. obtain such a bound assuming differential privacy and $0 \le \ell \le 1$.

Variations and Extensions

Tighter Bound via Individual-Sample Mutual Info

Theorem (Bu–Zou–Veeravalli). Suppose $\ell(w,Z)$ is σ^2-subgaussian for each $w \in \mathsf{W}$. Then for any learning algo. $P_{W|\boldsymbol{Z}}$ we have

$$|\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})| \le \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W;Z_i)}$$

This bound is tighter than the Xu–Raginsky bound:

$$\begin{aligned}
\sqrt{I(W;\boldsymbol{Z})} &= \sqrt{\sum_{i=1}^{n} I(W,Z^{i-1};Z_i)} && \text{(chain rule, independence)}\\
&\ge \sqrt{\sum_{i=1}^{n} I(W;Z_i)} && \text{(data processing)}\\
&\ge \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sqrt{I(W;Z_i)} && \text{(Jensen)}
\end{aligned}$$

Proof

1. Decompose

$$\begin{aligned}\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) &= \mathbf{E}[L_\mu(W) - L_{\boldsymbol{Z}}(W)] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[L_\mu(W) - \ell(W, Z_i)]\end{aligned}$$

2. Apply the Decoupling Estimate to each term in the sum: take $U = W$, $V = Z_i$, $f(U, V) = \ell(W, Z_i)$, then

$$|\mathbf{E}[L_\mu(W) - \ell(W, Z)]| \le \sqrt{2\sigma^2 I(W; Z_i)}$$

3. Triangle inequality:

$$\begin{aligned}|\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})| &= \left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[L_\mu(W) - \ell(W, Z_i)]\right| \\ &\le \frac{1}{n}\sum_{i=1}^{n} |\mathbf{E}[L_\mu(W) - \ell(W, Z_i)]| \\ &\le \frac{1}{n}\sum_{i=1}^{n} \sqrt{2\sigma^2 I(W; Z_i)}\end{aligned}$$

Tighter Bound via Random-Sample Mutual Info

Theorem. Under our usual assumptions,

$$|\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})| \le \sqrt{2\sigma^2 I(W; Z_J)},$$

where $J \sim \mathsf{Uniform}([n])$ is independent of $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ and W.

Remarks:

1. A learning algo. generalizes well if its output does not leak too much info. about any training instance chosen uniformly at random.
2. As we will see, this bound is tighter than our earlier bound involving $I(W; \boldsymbol{Z})$.

Preparations for the Proof

Lemma A. Let $Z_1, \ldots, Z_n$ be i.i.d. samples from μ, and let $J \sim \mathsf{Uniform}([n])$ be independent of the Z_i's. Then Z_J has distribution μ and is independent of J.

Proof. For any measurable $E \subseteq \mathsf{Z}$ and $j \in [n]$,

$$\begin{aligned} \mathbf{P}[Z_J \in E, J = j] &= \mathbf{P}[Z_J \in E | J = j]\mathbf{P}[J = j] \\ &= \frac{1}{n}\mathbf{P}[Z_j \in E] \\ &= \mathbf{P}[J = j] \cdot \mu(E) \end{aligned}$$

Marginalizing over J, we see that $\mathbf{P}[Z_J \in E] = \mu(E)$, so $Z_J \sim \mu$. Independence follows immediately.

Lemma B. For any $P_{W|\boldsymbol{Z}}$, Z_J and W are conditionally independent given $\boldsymbol{Z}$, and $\mathbf{E}[\ell(Z_J, W)|\boldsymbol{Z}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[\ell(Z_j, W)|\boldsymbol{Z}]$

Proof. Obvious.

Proof of the Random-Sample Mutual Info Bound

1. Let $\bar{W} \perp\!\!\!\perp (\boldsymbol{Z}, J)$ be an independent copy of W. Then, since $Z_J \sim \mu$,

$$\begin{aligned}
\mathbf{E}[L_\mu(W)] &= \mathbf{E}[\ell(\bar{W}, Z_J)] \\
\mathbf{E}[L_{\boldsymbol{Z}}(W)] &= \mathbf{E}\left[\mathbf{E}\left[\frac{1}{n}\sum_{j=1}^{n} \ell(W, Z_j)\middle|\boldsymbol{Z}\right]\right] \\
&= \mathbf{E}\left[\mathbf{E}[\ell(W, Z_J)|\boldsymbol{Z}]\right] \\
&= \mathbf{E}[\ell(W, Z_J)]
\end{aligned}$$

   Thus, $\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) = \mathbf{E}[\ell(\bar{W}, Z_J) - \ell(W, Z_J)]$.

2. Since $Z_J \sim \mu$, $\ell(w, Z_J)$ is σ^2-subgaussian for every $w \in \mathsf{W}$. Thus, by the Decoupling Estimate,

$$|\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})| \le \sqrt{2\sigma^2 I(W; Z_J)}.$$

Comparison with the Mutual Information Bound

$$\begin{aligned} I(W;Z_J) &\le I(W,J;Z_J) && \text{(data processing)} \\ &= I(W;Z_J|J) && \text{(chain rule, } Z_J \perp\!\!\!\perp J\text{)} \\ &= \frac{1}{n}\sum_{i=1}^{n} I(W;Z_i) \\ &\le \frac{1}{n} I(W;\boldsymbol{Z}) \end{aligned}$$

— the random-sample mutual information is not larger than $\frac{1}{n}I(W;\boldsymbol{Z})$

Part II

Information-Theoretically Stable Learning Algorithms

Information-Theoretically Stable Learning Algorithms

- From now on, assume $\ell(w, Z)$ is σ^2-subgaussian for all w.
- We will look at two commonly used types of learning algorithms and show that they are information-theoretically stable:
    1. The Gibbs algorithm.
    2. Iterative noisy algorithms (including Stochastic Gradient Langevin Dynamics).

Example 1: The Gibbs Algorithm

The Gibbs Algorithm

- Fix a data-independent *prior distribution* Q on W and a parameter $\beta > 0$.
- The *Gibbs algorithm* is given by

$$P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}}(\mathrm{d}w) = \frac{e^{-\beta L_{\boldsymbol{z}}(w)} Q(\mathrm{d}w)}{\mathbf{E}_Q[e^{-\beta L_{\boldsymbol{z}}(W)}]}$$

- The Gibbs algorithm is a 'soft' version of ERM: for any $\boldsymbol{z}$,

$$P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}} \xrightarrow{\beta \to \infty} P^{\mathsf{ERM}}_{W|\boldsymbol{Z}=\boldsymbol{z}} \qquad \text{in distribution}$$

Generalization bound for the Gibbs Algorithm

Theorem (Xu-Raginsky). Suppose that the loss function $\ell(w,z)$ takes values between 0 and 1. Then for any μ and any $\beta > 0$,

$$|\mathsf{gen}(\mu, P^{(\beta)}_{W|\boldsymbol{Z}})| \leq \frac{\beta}{2n}.$$

Proof

1. For any two $\boldsymbol{z}, \boldsymbol{z}'$ with $d_{\mathsf{H}}(\boldsymbol{z}, \boldsymbol{z}') = 1$ (say, $z_j = z_j'$ for all $j \neq i$),

$$L_{\boldsymbol{z}}(w) - L_{\boldsymbol{z}'}(w) = \frac{1}{n}\left[\ell(w, z_i) - \ell(w, z_i')\right] \in \left[-\frac{1}{n}, \frac{1}{n}\right]$$

2. We can use definitions and Hoeffding's lemma:

$$\begin{aligned}
D(P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}'}) &= \int_{\mathsf{W}} \mathrm{d}P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}} \log \frac{\mathrm{d}P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}'}} \\
&= \beta \int_{\mathsf{W}} \mathrm{d}P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}} \left(L_{\boldsymbol{z}'} - L_{\boldsymbol{z}}\right) + \log \frac{\mathbf{E}_Q[e^{-\beta L_{\boldsymbol{z}'}(W)}]}{\mathbf{E}_Q[e^{-\beta L_{\boldsymbol{z}}(W)}]} \\
&= \beta \int_{\mathsf{W}} \mathrm{d}P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}} \left(L_{\boldsymbol{z}'} - L_{\boldsymbol{z}}\right) + \log \int_{\mathsf{W}} \mathrm{d}P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}} e^{\beta(L_{\boldsymbol{z}} - L_{\boldsymbol{z}'})} \\
&\leq \underbrace{\beta \int_{\mathsf{W}} \mathrm{d}P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}} \left(L_{\boldsymbol{z}'} - L_{\boldsymbol{z}}\right) + \beta \int_{\mathsf{W}} \mathrm{d}P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}} \left(L_{\boldsymbol{z}} - L_{\boldsymbol{z}'}\right)}_{=0} + \frac{\beta^2}{2n^2}
\end{aligned}$$

3. Thus, for the Gibbs algorithm, $I(W; \boldsymbol{Z}) \leq I^-(W; \boldsymbol{Z}) \leq \frac{\beta^2}{2n}$

Gibbs Algorithm: An Origin Story

- Why this specific form for the Gibbs algorithm?
- Recall the information-theoretic bound for any $P_{W|\boldsymbol{Z}}$ (under subgaussianity):

$$\begin{aligned}\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) &= \mathbf{E}[L_\mu(W)] - \mathbf{E}[L_{\boldsymbol{Z}}(W)] \\ &\leq \sqrt{\frac{2\sigma^2}{n} I(W;\boldsymbol{Z})} \\ \implies \quad \mathbf{E}[L_\mu(W)] &\leq \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \sqrt{\frac{2\sigma^2}{n} I(W;\boldsymbol{Z})}\end{aligned}$$

Lemma. If $\ell(w, Z)$ is σ^2-subgaussian for any $w \in \mathsf{W}$, then, for any learning algo. $P_{W|\boldsymbol{Z}}$ and any $\beta > 0$

$$\mathbf{E}[L_\mu(W)] \leq \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \frac{I(W;\boldsymbol{Z})}{\beta} + \frac{\beta\sigma^2}{2n}$$

Lemma. If $\ell(w, Z)$ is σ^2-subgaussian for any $w \in \mathsf{W}$, then, for any learning algo. $P_{W|\boldsymbol{Z}}$ and any $\beta > 0$

$$\mathbf{E}[L_\mu(W)] \leq \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \frac{I(W; \boldsymbol{Z})}{\beta} + \frac{\beta\sigma^2}{2n}$$

Proof.

$$\begin{aligned}\mathbf{E}[L_\mu(W)] &\leq \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \sqrt{\frac{2\sigma^2}{n} I(W; \boldsymbol{Z})} \\ &= \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \inf_{\beta > 0} \left\{ \frac{I(W; \boldsymbol{Z})}{\beta} + \frac{\beta\sigma^2}{2n} \right\}\end{aligned}$$

The Gibbs Relaxation

Lemma. If $\ell(w, Z)$ is σ^2-subgaussian for any $w \in \mathsf{W}$, then, for any learning algo. $P_{W|\boldsymbol{Z}}$ and any $\beta > 0$

$$\mathbf{E}[L_\mu(W)] \le \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \frac{I(W; \boldsymbol{Z})}{\beta} + \frac{\beta\sigma^2}{2n}$$

- To construct $P_{W|\boldsymbol{Z}}$ with small $\mathbf{E}[L_\mu(W)]$, let's fix $\beta > 0$ and choose

$$P^*_{W|\boldsymbol{Z}} = \operatorname*{arg\,min}_{P_{W|\boldsymbol{Z}}} \left\{ \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \frac{I(W; \boldsymbol{Z})}{\beta} \right\}$$

- ***Bad idea!!!*** — $I(W; \boldsymbol{Z})$ depends on both $P_{W|\boldsymbol{Z}}$ and μ, but the latter is unknown! (That's why we need learning in the first place.)
- ***Better idea*** — introduce a relaxation.

The Gibbs relaxation

- Recall the Golden Formula: for any Q_W s.t. $D(P_W \| Q_W) < \infty$,

$$I(W; \boldsymbol{Z}) = D(P_{W|\boldsymbol{Z}} \| Q_W | P_{\boldsymbol{Z}}) - D(P_W \| Q_W)$$

- Consequently, for any $P_{W|\boldsymbol{Z}}$,

$$\begin{aligned}
&\mathbf{E}[L_{\boldsymbol{Z}}(W)] + \frac{I(W; \boldsymbol{Z})}{\beta} \\
&\le \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \frac{D(P_{W|\boldsymbol{Z}} \| Q_W | P_{\boldsymbol{Z}})}{\beta} \\
&= \int_{\mathsf{Z}^n} \mu^{\otimes n}(\mathrm{d}\boldsymbol{z}) \int_{\mathsf{W}} P_{W|\boldsymbol{Z}=\boldsymbol{z}}(\mathrm{d}w) \left\{ L_{\boldsymbol{z}}(w) + \frac{D(P_{W|\boldsymbol{Z}=\boldsymbol{z}} \| Q_W)}{\beta} \right\}
\end{aligned}$$

$$\begin{aligned}
&\min_{P_{W|\boldsymbol{Z}}} \left\{ \mathbf{E}[L_{\boldsymbol{Z}}(W)] + \frac{I(W; \boldsymbol{Z})}{\beta} \right\} \\
&\le \int_{\mathsf{Z}^n} \mu^{\otimes n}(\mathrm{d}\boldsymbol{z}) \min_{P_{W|\boldsymbol{Z}=\boldsymbol{z}}} \int_{\mathsf{W}} P_{W|\boldsymbol{Z}=\boldsymbol{z}}(\mathrm{d}w) \left\{ L_{\boldsymbol{z}}(w) + \frac{D(P_{W|\boldsymbol{Z}=\boldsymbol{z}} \| Q)}{\beta} \right\}
\end{aligned}$$

Enter the Gibbs Algorithm

Lemma. For any $Q \in \mathcal{P}(\mathsf{W})$,

$$\int_{\mathsf{W}} P_{W|\boldsymbol{Z}=\boldsymbol{z}}(\mathrm{d}w) \left\{ L_{\boldsymbol{z}}(w) + \frac{D(P_{W|\boldsymbol{Z}=\boldsymbol{z}} \| Q)}{\beta} \right\} \ge -\frac{1}{\beta} \log \mathbf{E}_Q[e^{-\beta L_{\boldsymbol{z}}(W)}],$$

and the minimum is achieved uniquely by the Gibbs algorithm

$$P^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}}(\mathrm{d}w) = \frac{e^{-\beta L_{\boldsymbol{z}}(w)} Q(\mathrm{d}w)}{\mathbf{E}_Q[e^{-\beta L_{\boldsymbol{z}}(W)}]}$$

Proof. Exercise.

Expected Risk Bound for the Gibbs Algorithm

Putting everything together ...

Theorem. If $\ell(w, Z)$ is σ-subgaussian for any $w \in \mathsf{W}$, then the expected risk of the Gibbs algorithm $P^{(\beta)}_{W|\boldsymbol{Z}}$ satisfies

$$\mathbf{E}[L_\mu(W)] \le -\frac{1}{\beta}\mathbf{E}\left\{\log \mathbf{E}[e^{-\beta L_{\boldsymbol{Z}}(\bar{W})}|\boldsymbol{Z}]\right\} + \frac{\beta\sigma^2}{2n},$$

where $\bar{W} \sim Q$ and $\boldsymbol{Z} \sim \mu^{\otimes n}$ are independent.

Remarks:

1. Bounds of this sort are known as *PAC-Bayesian bounds* (Catoni; McAllester; Ortiz; Zhang ...)
2. Now the whole affair hinges on being able to bound the log-partition function $\mathbf{E}_Q[e^{-\beta L_{\boldsymbol{z}}(\bar{W})}]$, uniformly in $\boldsymbol{z}$.

Example: Smooth Losses

- Consider the case of $\mathsf{W} = \mathbb{R}^d$, $\ell(w, z)$ differentiable in w and *L-smooth* (∇ w.r.t. the first argument):

$$\|\nabla \ell(w, z) - \nabla \ell(v, z)\| \le L\|w - v\|$$

- We are *not* assuming that $w \mapsto \ell(w, z)$ is convex.
- Let's choose the Gaussian prior

$$Q(\mathrm{d}w) = \frac{1}{(2\pi\rho^2)^{d/2}} \exp\left(-\frac{\|w\|^2}{2\rho^2}\right) \mathrm{d}w,$$

  where $\rho^2 > 0$ is a tunable parameter.

- Gibbs algorithm: given data $\boldsymbol{z}$, draw W from the density

$$p^{(\beta)}_{W|\boldsymbol{Z}=\boldsymbol{z}}(w) \propto \exp\left\{-\left(\frac{\beta}{n}\sum_{i=1}^{n} \ell(w, z_i) + \frac{1}{2\rho^2}\|w\|^2\right)\right\}$$

Excess Risk of Gibbs with Smooth Losses

Theorem. Assume the following:

1. $\ell(w, Z)$ is σ^2-subgaussian for every $w \in \mathsf{W}$.
2. $\ell(w, z)$ is differentiable in w, and
$$\sup_{z \in \mathsf{Z}} \|\nabla \ell(w, z) - \nabla \ell(v, z)\| \leq L\|w - v\|.$$
3. For all $\boldsymbol{z} \in \mathsf{Z}^n$, all global minimizers of $L_{\boldsymbol{z}}(w)$ lie in the ball of radius R.

Then for the Gibbs algo. $P^{(\beta)}_{W|\boldsymbol{Z}}$ with Gaussian prior $Q = \mathcal{N}(0, \rho^2 I_d)$

$$\mathbf{E}[L_\mu(W)] \leq \min_w L_\mu(w) + \frac{L\pi\rho^2 d}{\beta} + \frac{1}{2\beta\rho^2}\left(R + \sqrt{\frac{2\pi\rho^2 d}{\beta}}\right)^2 + \frac{d}{2\beta}\log\left(\frac{\beta}{d}\right) - \frac{1}{\beta}\log \mathcal{V}_d + \frac{\beta\sigma^2}{2n}$$

where $\mathcal{V}_d$ is the volume of the unit ball in $(\mathbb{R}^d, \|\cdot\|)$.

Proof

1. Fix $\boldsymbol{z}$, let $w^*_{\boldsymbol{z}}$ be any global minimizer of $L_{\boldsymbol{z}}(w)$; recall $\|w^*_{\boldsymbol{z}}\| \le R$.
2. Since $\ell(w, z)$ is L-smooth,

$$L_{\boldsymbol{z}}(w) - L_{\boldsymbol{z}}(w^*_{\boldsymbol{z}}) \le \frac{L}{2}\|w - w^*_{\boldsymbol{z}}\|^2$$

   (see any text on optimization, e.g., Nesterov).

3. Now we can estimate (recall $Q = \mathcal{N}(0, \rho^2 I_d)$):

$$\begin{aligned}\int_{\mathbb{R}^d} Q(\mathrm{d}w) e^{-\beta L_{\boldsymbol{z}}(w)} &= e^{-\beta L_{\boldsymbol{z}}(w^*_{\boldsymbol{z}})} \int_{\mathbb{R}^d} Q(\mathrm{d}w) e^{-\beta(L_{\boldsymbol{z}}(w) - L_{\boldsymbol{z}}(w^*_{\boldsymbol{z}}))} \\ &\ge e^{-\beta L_{\boldsymbol{z}}(w^*_{\boldsymbol{z}})} \int_{\mathbb{R}^d} Q(\mathrm{d}w) e^{-\frac{\beta L}{2}\|w - w^*_{\boldsymbol{z}}\|^2}\end{aligned}$$

4. It remains to lower-bound the Gaussian integral

$$G = \frac{1}{(2\pi\rho^2)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{1}{2\rho^2}\|w\|^2} e^{-\frac{\beta L}{2}\|w - w^*_{\boldsymbol{z}}\|^2}\,\mathrm{d}w$$

Proof (cont'd)

4. We want to lower-bound the Gaussian integral

$$G = \frac{1}{(2\pi\rho^2)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{1}{2\rho^2}\|w\|^2} e^{-\frac{\beta L}{2}\|w - w^*_{\boldsymbol{z}}\|^2} \, \mathrm{d}w$$

5. Let $\mathcal{B} := B_2^d(w^*_{\boldsymbol{z}}, \varepsilon)$, radius $\varepsilon > 0$ to be tuned later. Then:

$$\begin{aligned} G &\geq \frac{1}{(2\pi\rho^2)^{d/2}} e^{-\frac{\beta L \varepsilon^2}{2}} \cdot \int_{\mathcal{B}} e^{-\frac{1}{2\rho^2}\|w\|^2} \, \mathrm{d}w \\ &\geq \frac{1}{(2\pi\rho^2)^{d/2}} e^{-\frac{\beta L \varepsilon^2}{2}} \cdot e^{-\frac{1}{2\rho^2}(\|w^*_{\boldsymbol{z}}\| + \varepsilon)^2} \mathsf{Vol}_d(\mathcal{B}) \\ &= \frac{1}{(2\pi\rho^2)^{d/2}} e^{-\frac{\beta L \varepsilon^2}{2}} \cdot e^{-\frac{1}{2\rho^2}(\|w^*_{\boldsymbol{z}}\| + \varepsilon)^2} \varepsilon^d \mathcal{V}_d \\ &\geq \left(\frac{\varepsilon^2}{2\pi\rho^2} \right)^{d/2} \exp\left(-\frac{\beta L \varepsilon^2}{2} - \frac{1}{2\rho^2}(R + \varepsilon)^2 \right) \mathcal{V}_d \end{aligned}$$

Proof (Almost There ...)

6. Now we can estimate

$$
\begin{aligned}
-\frac{1}{\beta}\mathbf{E}\left\{\log \mathbf{E}[e^{-\beta L_{\boldsymbol{Z}}(\bar{W})}|\boldsymbol{Z}]\right\} &\le \mathbf{E}\left[\min_{w\in\mathsf{W}} L_{\boldsymbol{Z}}(w)\right] \\
&\quad + \frac{L\varepsilon^2}{2} + \frac{1}{2\beta\rho^2}(R+\varepsilon)^2 + \frac{d}{2\beta}\log\left(\frac{2\pi\rho^2}{\varepsilon^2}\right) - \frac{1}{\beta}\log\mathcal{V}_d, \quad \forall\varepsilon > 0
\end{aligned}
$$

7. Choose $\varepsilon^2 = \frac{2\pi\rho^2 d}{\beta}$ to get

$$
\begin{aligned}
(\cdots) &\le \mathbf{E}\left[\min_{w} L_{\boldsymbol{Z}}(w)\right] \\
&\quad + \frac{L\pi\rho^2 d}{\beta} + \frac{1}{2\beta\rho^2}\left(R + \sqrt{\frac{2\pi\rho^2 d}{\beta}}\right)^2 + \frac{d}{2\beta}\log\left(\frac{\beta}{d}\right) - \frac{1}{\beta}\log\mathcal{V}_d
\end{aligned}
$$

8. Finally, let w^* be any minimizer of $L_\mu(w)$ and note that

$$
\mathbf{E}\left[\min_{w} L_{\boldsymbol{Z}}(w)\right] \le \mathbf{E}[L_{\boldsymbol{Z}}(w^*)] = \min_{w} L_\mu(w).
$$

Example 2: Noisy Iterative Algorithms

The Set-Up

- Suppose $\mathsf{W} = \mathbb{R}^d$, as before.
- We generate W as follows:

$$
\begin{aligned}
&W = f(V_1, \ldots, V_T) \\
&V_0 \text{ chosen randomly, independently of everything else} \\
&V_t = g(V_{t-1}) - \eta_t F(V_{t-1}, Z_{J_t}) + \xi_t, \qquad t = 1, 2, \ldots, T
\end{aligned}
$$

where:

- $T \in \mathbb{N}$ is a fixed number of iterations
- $J_1, \ldots, J_t$ is a sequence of random elements of $[n]$
- $\{\eta_t\}_{t=1}^T$ is a sequence of positive step sizes
- $\xi_t \sim \mathcal{N}(0, \rho_t^2 I_d)$ is a sequence of independent Gaussians, independent of everything else
- $f(\cdot)$, $g(\cdot)$ and $F(\cdot, \cdot)$ are deterministic mappings

Pensia, Jog, Loh (2018)

Example: Stochastic Gradient Langevin Dynamics

Assume $\ell(w, z)$ is differentiable in w

$$\begin{aligned} V_0 &= 0 \\ V_t &= V_{t-1} - \eta_t \nabla \ell(V_{t-1}, Z_{J_t}) + \xi_t, \qquad t = 1, \dots, T \\ W &= V_T \end{aligned}$$

where:

- $J_1, \dots, J_T \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}([n])$
- $\{\eta_t\}_{t=1}^T$ are positive step sizes
- $\xi_t \sim \mathcal{N}(0, \rho_t^2 I_d)$ with $\rho_t = \frac{\eta_t}{\beta}$ for some $\beta > 0$
- $g(v) = v$, $F(v, z) = \nabla \ell(v, z)$, $f(v_1, \dots, v_T) = v_T$

Other choices of f are possible, e.g., $f(v_1, \dots, v_T) = \frac{1}{T} \sum_{t=1}^T v_t$ (trajectory averaging) ...

Assumptions and Goal

- *Sampling strategy*: conditional distribution of $\boldsymbol{J} = (J_1, \ldots, J_T)$ given $(\boldsymbol{Z}, \boldsymbol{V})$
- Assumption 1: the sampling strategy is such that, for every $t \in [T-1]$,

$$P_{J_{t+1}|J_1,\ldots,J_t,\boldsymbol{V},\boldsymbol{Z}} = P_{J_{t+1}|J_1,\ldots,J_t,\boldsymbol{Z}}$$

  — that is, the index of the sample in the next round does not depend on the iterates $V_1, \ldots, V_t$, given the previous choices $J_1, \ldots, J_t$ and data $\boldsymbol{Z}$
- Assumption 2: the update function $F(\cdot, \cdot)$ is bounded:

$$\sup_{v \in \mathbb{R}^d} \sup_{z \in \mathsf{Z}} \|F(v, z)\| \leq L < \infty$$

- To control the generalization error, we will upper-bound the mutual information $I(W; \boldsymbol{Z})$

Mutual Information $I(W;\boldsymbol{Z})$

- Let $\boldsymbol{Z}^{\boldsymbol{J}} := (Z_{J_1}, \ldots, Z_{J_T})$ and note that

$$\boldsymbol{Z} \longleftrightarrow \boldsymbol{Z}^{\boldsymbol{J}} \longleftrightarrow \boldsymbol{V}$$

- Then

$$\begin{aligned} I(W;\boldsymbol{Z}) &= I(f(\boldsymbol{V});\boldsymbol{Z}) \\ &\leq I(\boldsymbol{V};\boldsymbol{Z}) && \text{(data processing)} \\ &\leq I(\boldsymbol{V};\boldsymbol{Z}^{\boldsymbol{J}}) && \text{(data processing again)} \\ &= \sum_{t=1}^{T} I(V_t;\boldsymbol{Z}^{\boldsymbol{J}}|V^{t-1}) && \text{(chain rule)} \end{aligned}$$

  so now we will analyze each of the conditional mutual information terms

- By definition,

$$I(V_t;\boldsymbol{Z}^{\boldsymbol{J}}|V^{t-1}) = h(V_t|V^{t-1}) - h(V_t|V^{t-1},\boldsymbol{Z}^{\boldsymbol{J}})$$

  where $h(\cdot)$ is the *differential entropy*

Conditional Mutual Information $I(V_t; \boldsymbol{Z}^{\boldsymbol{J}} | V^{t-1})$

- Recall the stochastic update

$$V_t = g(V_{t-1}) - \eta_t F(V_{t-1}, Z_{J_t}) + \xi_t$$

which implies that $V_t \longleftrightarrow (V_{t-1}, Z_{J_t}) \longleftrightarrow (V^{t-2}, \boldsymbol{Z}^{\boldsymbol{J} \setminus \{J_t\}})$

- Thus:

$$\begin{aligned} h(V_t | V^{t-1}, \boldsymbol{Z}^{\boldsymbol{J}}) &= h(V_t | V_{t-1}, Z_{J_t}, V^{t-2}, \boldsymbol{Z}^{\boldsymbol{J} \setminus \{J_t\}}) \\ &= h(V_t | V_{t-1}, Z_{J_t}) \end{aligned}$$

- By the same token,

$$h(V_t | V^{t-1}) = h(V_t | V_{t-1})$$

Lemma (Pensia–Jog–Loh). Under the conditional independence assumption on the sampling strategy,

$$I(V_t; \boldsymbol{Z}^{\boldsymbol{J}} | V^{t-1}) = h(V_t | V_{t-1}) - h(V_t | V_{t-1}, Z_{J_t}) = I(V_t; Z_{J_t} | V_{t-1})$$

Conditional Mutual Information $I(V_t; Z_{J_t}|V_{t-1})$

- Conditionally on $V_{t-1} = v_{t-1}$,

$$V_t = g(v_{t-1}) - \eta_t F(v_{t-1}, Z_{J_t}) + \xi_t, \qquad Z_{J_t} \perp\!\!\!\perp \xi_t$$

- Then, by shift-invariance of differential entropy,

$$\begin{aligned} h(V_t|V_{t-1} = v_{t-1}) &= h(V_t - g(v_{t-1})|V_{t-1} = v_{t-1}) \\ &= h(-\eta_t F(v_{t-1}, Z_{J-t}) + \xi_t|V_{t-1} = v_{t-1}) \end{aligned}$$

- Recall: for any d-dim. random vector U with $\mathbf{E}\|U\|^2 < \infty$,

$$h(U) \le \frac{d}{2} \log\left(\frac{2\pi e \mathbf{E}\|U\|^2}{d}\right)$$

- Since $Z_{J_t} \perp\!\!\!\perp \xi_t$ and $\mathbf{E}[\xi_t] = 0$,

$$\begin{aligned} &\mathbf{E}[\| - \eta_t F(v_{t-1}, Z_{J_t}) + \xi_t\|^2|V_{t-1} = v_{t-1}] \\ &\qquad = \eta_t^2 \mathbf{E}[\|F(v_{t-1}, Z_{J_t})\|^2|V_{t-1}] + \mathbf{E}\|\xi_t\|^2 \le \eta_t^2 L^2 + d\rho_t^2 \end{aligned}$$

$$\Longrightarrow h(V_t|V_{t-1}) \le \frac{d}{2} \log\left(\frac{2\pi e(\eta_t^2 L^2 + d\rho_t^2)}{d}\right)$$

Conditional Mutual Information $I(V_t; Z_{J_t} | V_{t-1})$

- By the same reasoning,

$$\begin{aligned} & h(V_t | V_{t-1}, Z_{J_t}) \\ & = h(g(V_{t-1}) - \eta_t F(V_{t-1}, Z_{J_t}) + \xi_t | V_{t-1}, Z_{J_t}) \\ & = h(\xi_t | V_{t-1}, Z_{J_t}) \\ & = h(\xi_t) && \xi_t \perp\!\!\!\perp (V_t, Z_{J_t}) \\ & = \frac{d}{2} \log(2\pi e \rho_t^2) && \xi_t \sim \mathcal{N}(0, \rho_t^2 I_d) \end{aligned}$$

Lemma (Pensia–Jog–Loh). For every $t \in [T]$,

$$I(V_t; Z_{J_t} | V_{t-1}) \leq \frac{d}{2} \log\left(1 + \frac{\eta_t^2 L^2}{d\rho_t^2}\right) \leq \frac{\eta_t^2 L^2}{2\rho_t^2}$$

Generalization Bound for Noisy, Iteratie Algorithms

Recall the processing pipeline:

$$\begin{aligned}
&\boldsymbol{Z} \longrightarrow (V_1, \ldots, V_T) \longrightarrow W \\
&V_t = g(V_{t-1}) - \eta_t F(V_{t-1}, Z_{J_t}) + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \rho_t^2 I_d) \\
&W = f(V_1, \ldots, V_T)
\end{aligned}$$

where we assume:

1. J_t is conditionally independent of $\boldsymbol{V}$ given $(J_1, \ldots, J_{t-1})$ and $\boldsymbol{Z}$
2. $\|F(\cdot, \cdot)\| \leq L$
3. $\ell(w, Z)$ is σ^2-subgaussian for every w

Theorem (Pensia–Jog–Loh). Under the above assumptions,

$$\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}}) \leq \sqrt{\frac{\sigma^2}{n} \sum_{t=1}^{T} \frac{\eta_t^2 L^2}{\rho_t^2}}$$

Generalization Bound for SGLD

- Assume $\ell(w,z)$ is differentiable in w and L-Lipschitz:

$$\sup_{z \in \mathsf{Z}} |\ell(w,z) - \ell(v,z)| \le L\|w - v\|$$

- Generate $V_1, \dots, V_T$: $V_0 = 0$ (say)

$$V_t = V_{t-1} - \eta_t \nabla \ell(V_{t-1}, Z_{J_t}) + \sqrt{\frac{\eta_t}{\beta}}\bar{\xi}_t, \qquad \bar{\xi}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$$

$$W = f(V_1, \dots, V_T) \qquad \text{— arbitrary postprocessing}$$

where $T = nk$ for some k and $\eta_t = \frac{1}{t}$

Theorem (Pensia–Jog–Loh).

$$|\mathsf{gen}(\mu, P_{W|\boldsymbol{Z}})| \le \sqrt{\frac{\beta\sigma^2 L^2}{n} \sum_{t=1}^{nk} \frac{1}{t}} \le \sqrt{\frac{\beta\sigma^2 L^2}{n}(\log n + \log k + 1)}$$

What I (Unfortunately) Had to Skip

- Preservation of stability under *adaptive composition*:

  $$P_{W_1,\ldots,W_k|\boldsymbol{Z}} = P_{W_1|\boldsymbol{Z}} \otimes P_{W_2|W_1,\boldsymbol{Z}} \otimes \ldots \otimes P_{W_k|W_1,\ldots,W_{k-1},\boldsymbol{Z}}$$

  — need to require that $I^{-}(W;\boldsymbol{Z}) \le n\varepsilon$ for *any* (not necessarily product) distribution of $\boldsymbol{Z}$ (Steinke–Feldman, 2018)
- Refined bounds for Gibbs-type and other differentially private algorithms (Wang–Lei–Fienberg, 2016; Dziugaite–Roy, 2019); Kuzborskij–Cesa-Bianchi–Szepesvári, 2019)
- Other notions of information: max-information (Dwork et al., 2016); Rényi information and divergence (Mironov, 2017); concentrated differential privacy (Dwork and Rothblum, 2016; Bun–Steinke, 2016)
- Total-variation and Wasserstein stability (Raginsky–Rakhlin–Tsao–Wu–Xu, 2016; Alabdulmohsin, 2017; Lopez–Jog, 2018)
- Refined bounds via mutual information and chaining for subgaussian processes (Asadi–Abbe–Verdú, 2018; Asadi–Abbe, 2019)

Some Open Problems

1. Can we get ‘high-probability’ bounds on the generalization error for information-theoretically stable learning algorithms? Perhaps, under additional assumptions?
2. Can we prove information-theoretic stability of Stochastic Gradient Descent (SGD), without additional noisy perturbations? What about *deterministic* learning algorithms? Some results by Raginsky–Rakhlin–Tsao–Wu–Xu, 2016; Bu–Zou–Veeravalli, 2019
3. Converse results? Does poor generalization imply information leakage? Some results by Bassily et al., 2018; Nachum–Shafer–Yehudayoff, 2018; Nachum–Yehudayoff, 2019