Generalization Bounds via Convex Analysis

Gergely Neu

Universitat Pompeu Fabra Barcelona

joint work with Gábor Lugosi

Funded by
ERC StG
ScaleR

Outline

- Supervised learning crash course
- Beyond “information theoretic” generalization
- Generalization bounds via convex analysis
- Classic examples: relative entropy, χ^2, p-norm…
- New (and cool?) example: smoothed relative entropy
- Some words about the proof

Outline

- Supervised learning crash course
- Beyond “information theoretic” generalization
- Generalization bounds via convex analysis
- Classic examples: relative entropy, χ^2, p-norm…
- New (and cool?) example: smoothed relative entropy
- Some words about the proof

Setup: Supervised learning

- Data set: $S_n = \{Z_i\}_{i=1}^n \in \mathcal{Z}^n = \mathcal{S}$, drawn i.i.d. $\sim\mu$
  - e.g., regression: $Z_i = (X_i, Y_i)$ with $X_i \in \mathbb{R}^m$ and $Y_i \in \mathbb{R}$
- Hypothesis class: $\mathcal{W}$
  - e.g., neural network weights
- Loss function: $\ell: \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$
  - e.g., square loss: $\ell\big(w, (x, y)\big) = (f(w, x) - y)^2$
- Learning algorithm $\mathcal{A}: \mathcal{S} \to \mathcal{W}$ produces hypothesis $W_n = \mathcal{A}(S_n)$

Setup: Supervised learning

- Data set: $S_n = \{Z_i\}_{i=1}^n \in \mathcal{Z}^n = \mathcal{S}$, drawn i.i.d. $\sim\mu$
  - e.g., regression: $Z_i = (X_i, Y_i)$ with $X_i \in \mathbb{R}^m$ and $Y_i \in \mathbb{R}$
- Hypothesis class: $\mathcal{W}$
  - e.g., neural network weights
- Loss function: $\ell: \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$
  - e.g., square loss: $\ell\big(w, (x, y)\big) = (f(w, x) - y)^2$
- Learning algorithm $\mathcal{A}: \mathcal{S} \to \mathcal{W}$ produces hypothesis $W_n = \mathcal{A}(S_n)$

Generalization error:

$$\mathrm{gen}(W_n, S_n) = \frac{1}{n}\sum_{i=1}^n(\ell(W_n, Z_i) - \mathbb{E}[\ell(W_n, Z')|W_n])$$

Information-theoretic generalization

Theorem

(Russo & Zou, 2016, Xu & Raginsky, 2017)

Suppose that $\ell(w, Z)$ is σ-subgaussian for all $w \in \mathcal{W}$.

Then, for any learning algorithm $\mathcal{A}$,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]| \leq \sqrt{\frac{2\sigma^2 \mathcal{D}_{\mathrm{KL}}\left(P_{W_n,S_n} \middle| P_{W_n} \otimes P_{S_n}\right)}{n}}$$

Information-theoretic generalization

Theorem

(Russo & Zou, 2016, Xu & Raginsky, 2017)

Suppose that $\ell(w, Z)$ is σ-subgaussian for all $w \in \mathcal{W}$.
Then, for any learning algorithm $\mathcal{A}$,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]| \leq \sqrt{\frac{2\sigma^2 \mathcal{D}_{\mathrm{KL}}\left(P_{W_n,S_n} | P_{W_n} \otimes P_{S_n}\right)}{n}}$$

What's special about $\mathcal{D}_{\mathrm{KL}}$?

More concretely:

Can we replace $\mathcal{D}_{\mathrm{KL}}$ by another function H and get

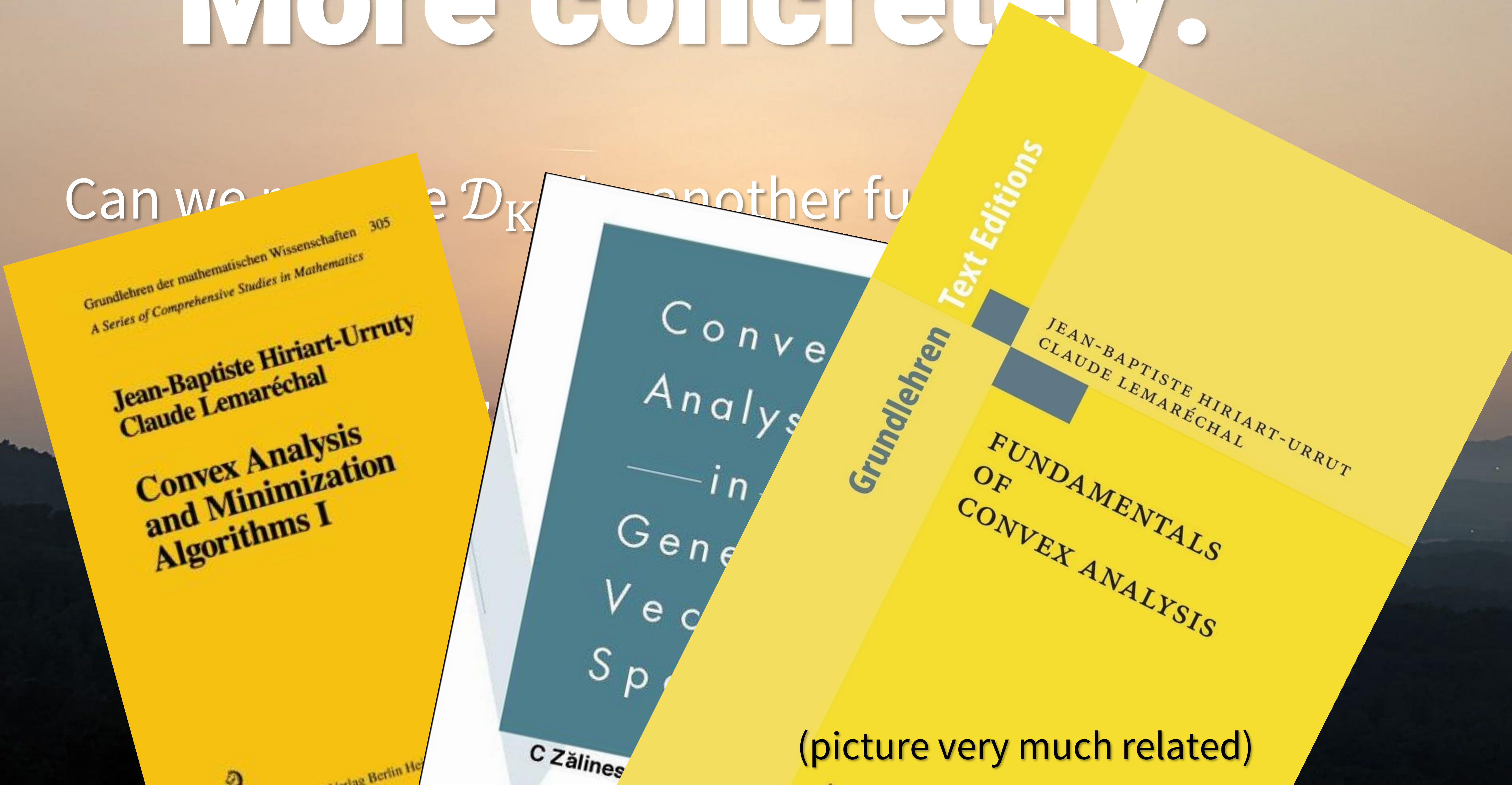
$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]| \leq \sqrt{\frac{\mathrm{const} \cdot H\left(P_{W_n, S_n}\right)}{n}}$$

?

(picture unrelated)

More concretely:

Can we r e $\mathcal{D}_{\mathrm{K}}$ another fu



(picture very much related)

Outline

- Supervised learning crash course
- Beyond “information theoretic” generalization
- Generalization bounds via convex analysis
- Classic examples: relative entropy, χ^2, p-norm…
- New (and cool?) example: smoothed relative entropy
- Some words about the proof

Notation

- Δ ={distributions P on $\mathcal{W} \times \mathcal{S}$ with $\mathcal{S}$ -marginal $\mu^{\otimes n}$}
  - Important special choices: $P_n = P_{W_n,S_n}$ and $\overline{P}_n = P_{W_n} \otimes P_{S_n}$
- $\mathcal{F}$ ={bounded measurable functions $f\colon \mathcal{W} \times \mathcal{S} \to \mathbb{R}$}

Notation

- Δ ={distributions P on $\mathcal{W} \times \mathcal{S}$ with $\mathcal{S}$ -marginal $\mu^{\otimes n}$}
  - Important special choices: $P_n = P_{W_n,S_n}$ and $\overline{P}_n = P_{W_n} \otimes P_{S_n}$
- $\mathcal{F}$ ={bounded measurable functions $f: \mathcal{W} \times \mathcal{S} \to \mathbb{R}$}
- For any $P \in \Delta$ and $f \in \mathcal{F}$, define

$$\langle P, f \rangle = \mathbb{E}_{(W,S)\sim P}[f(W,S)]$$

Notation

- Δ ={distributions P on $\mathcal{W} \times \mathcal{S}$ with $\mathcal{S}$ -marginal $\mu^{\otimes n}$}
  - Important special choices: $P_n = P_{W_n,S_n}$ and $\overline{P}_n = P_{W_n} \otimes P_{S_n}$
- $\mathcal{F}$ ={bounded measurable functions $f: \mathcal{W} \times \mathcal{S} \to \mathbb{R}$}
- For any $P \in \Delta$ and $f \in \mathcal{F}$, define
$$\langle P, f \rangle = \mathbb{E}_{(W,S)\sim P}[f(W,S)]$$
- Centered loss: $\overline{\ell}(w,z) = \ell(w,z) - \mathbb{E}[\ell(w,Z')]$
- Centered average loss: $\overline{L}_n(w,s) = \frac{1}{n}\sum_{i=1}^{n} \overline{\ell}\,(w,z_i)$

Notation

- Δ ={distributions P on $\mathcal{W} \times \mathcal{S}$ with $\mathcal{S}$ -marginal $\mu^{\otimes n}$}
  - Important special choices: $P_n = P_{W_n,S_n}$ and $\overline{P}_n = P_{W_n} \otimes P_{S_n}$
- $\mathcal{F}$ ={bounded measurable functions $f: \mathcal{W} \times \mathcal{S} \to \mathbb{R}$}
- For any $P \in \Delta$ and $f \in \mathcal{F}$, define

$$\langle P, f \rangle = \mathbb{E}_{(W,S)\sim P}[f(W,S)]$$

- Centered loss: $\overline{\ell}(w,z) = \ell(w,z) - \mathbb{E}[\ell(w,Z')]$
- Centered average loss: $\overline{L}_n(w,s) = \frac{1}{n}\sum_{i=1}^{n} \overline{\ell}\,(w,z_i)$

Expected generalization error:

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)] = \langle P_n, \overline{L}_n \rangle$$

Notation++

- Let $H : \Delta \to \mathbb{R}_+$ be convex: $\forall P, P' \in \Delta, \lambda \in [0,1]$:
$$H(\lambda P + (1-\lambda)P') \leq \lambda H(P) + (1-\lambda)H(P')$$
- Legendre–Fenchel conjugate of H defined for all $f \in \mathcal{F}$ as
$$H^*(f) = \sup_{P \in \Delta}\{\langle P, f\rangle - H(P)\}$$
- Fenchel–Young inequality: for any $P \in \Delta$ and $f \in \mathcal{F}$,
$$\langle P, f\rangle \leq H(P) + H^*(f)$$

A generalization bound

For any $\eta \in \mathbb{R}$:

$$\eta\langle P_{W_n,S_n}, \overline{L}_n\rangle \leq H\left(P_{W_n,S_n}\right) + H^*\left(\eta\overline{L}_n\right)$$

A generalization bound

For any $\eta \in \mathbb{R}$:

$$\eta\langle P_{W_n,S_n}, \bar{L}_n\rangle \le H\left(P_{W_n,S_n}\right) + H^*\left(\eta \bar{L}_n\right)$$

When can this be $O(\eta^2/n)$?

(that would imply a bound of order $\frac{H(P_n)}{\eta} + \frac{C\eta}{n} \sim \sqrt{\frac{CH(P_n)}{n}}$)

Example 1: relative entropy

- $H(P) = \mathcal{D}_{\mathrm{KL}}\left(P \middle| \overline{P}_n\right)$
- Conjugate: $H^*(f) = \log \mathbb{E}_{W_n, S_n'}\left[\exp\left(f(W_n, S_n')\right)\right]$ (Donsker-Varadhan formula)
- Applied to $\eta \overline{L}_n$:

$$H^*\left(\eta \overline{L}_n\right) = \log \mathbb{E}_{W_n, S_n'}\left[\exp\left(\frac{\eta}{n} \sum_{i=1}^{n} \overline{\ell}(W_n, Z_i')\right)\right] \leq \frac{\eta^2 \sigma^2}{n}$$

Example 1: relative entropy

- $H(P) = \mathcal{D}_{\mathrm{KL}}\left(P \middle| \overline{P}_n\right)$
- Conjugate: $H^*(f) = \log \mathbb{E}_{W_n, S_n'}\left[\exp\left(f(W_n, S_n')\right)\right]$ (Donsker-Varadhan formula)
- Applied to $\eta \overline{L}_n$:

$$H^*(\eta \overline{L}_n) = \log \mathbb{E}_{W_n, S_n'}\left[\exp\left(\frac{\eta}{n} \sum_{i=1}^{n} \overline{\ell}(W_n, Z_i')\right)\right] \leq \frac{\eta^2 \sigma^2}{n}$$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] \leq \frac{\mathcal{D}_{\mathrm{KL}}\left(P_n \middle| \overline{P}_n\right)}{\eta} + \frac{\eta \sigma^2}{n} \sim \sqrt{\frac{\sigma^2 \mathcal{D}_{\mathrm{KL}}\left(P_n \middle| \overline{P}_n\right)}{n}}$$

Example 2: χ^2-divergence

- $H(P) = \mathcal{D}_{\chi^2}\left(P \middle| \overline{P}_n\right) = \int \frac{\left(\mathrm{d}P - \mathrm{d}\overline{P}_n\right)^2}{\mathrm{d}\overline{P}_n}$
- Conjugate: $H^*(f) = \mathbb{E}_{W_n, S_n'}[(f(W_n, S_n') - \mathbb{E}[f(W_n, S_n')])^2]$
- Applied to $\eta \overline{L}_n$:

$$H^*\left(\eta \overline{L}_n\right) = \mathbb{E}_{W_n, S_n'}\left[\left(\frac{\eta}{n} \sum_{i=1}^{n}\left(\overline{\ell}(W_n, Z_i') - \mathbb{E}\left[\overline{\ell}(W_n, Z_i')\right]\right)\right)^2\right] = \frac{\eta^2 \mathrm{Var}\left[\overline{\ell}(W_n, Z')\right]}{n}$$

Example 2: χ^2-divergence

- $H(P) = \mathcal{D}_{\chi^2}(P|\overline{P}_n) = \int \frac{(\mathrm{d}P - \mathrm{d}\overline{P}_n)^2}{\mathrm{d}\overline{P}_n}$
- Conjugate: $H^*(f) = \mathbb{E}_{W_n,S_n'}[(f(W_n, S_n') - \mathbb{E}[f(W_n, S_n')])^2]$
- Applied to $\eta \overline{L}_n$:

$$H^*(\eta \overline{L}_n) = \mathbb{E}_{W_n,S_n'}\left[\left(\frac{\eta}{n}\sum_{i=1}^n \left(\overline{\ell}(W_n, Z_i') - \mathbb{E}\left[\overline{\ell}(W_n, Z_i')\right]\right)\right)^2\right] = \frac{\eta^2 \mathrm{Var}[\overline{\ell}(W_n,Z')]}{n}$$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] \le \frac{\mathcal{D}_{\chi^2}(P_n|\overline{P}_n)}{\eta} + \frac{\eta \mathrm{Var}\left[\overline{\ell}(W_n, Z')\right]}{n} \sim \sqrt{\frac{\mathrm{Var}\left[\overline{\ell}(W_n, Z')\right]\mathcal{D}_{\chi^2}(P_n|\overline{P}_n)}{n}}$$

(Is this known?)

Seriously, how do we pick H?

Seriously, how do we pick H?

- We will consider functions H of the form
$$H(P) = \mathbb{E}_S\left[h\left(P_{|S}\right)\right],$$
where
  - $P_{|s}$ is the conditional distribution of $W|S = s$ under $(W, S) \sim P$
  - h is a convex function acting on distributions over $\mathcal{W}$:
$$\forall Q, Q' \in \mathrm{Dist}(\mathcal{W}), \lambda \in [0,1]:$$
$$h(\lambda Q + (1-\lambda)Q') \leq \lambda h(Q) + (1-\lambda)h(Q')$$

Seriously, how do we pick H?

- We will consider functions H of the form
$$H(P) = \mathbb{E}_S\left[h\left(P_{|S}\right)\right],$$
where
  - $P_{|s}$ is the conditional distribution of $W|S = s$ under $(W, S) \sim P$
  - h is a convex function acting on distributions over $\mathcal{W}$:
$$\forall Q, Q' \in \mathrm{Dist}(\mathcal{W}), \lambda \in [0,1]:$$
$$h(\lambda Q + (1-\lambda)Q') \leq \lambda h(Q) + (1-\lambda)h(Q')$$
  - h is α-strongly convex wrt some norm $\|\cdot\|^2$:
$$\forall Q, Q' \in \mathrm{Dist}(\mathcal{W}), \lambda \in [0,1]:$$
$$h(\lambda Q + (1-\lambda)Q') \leq \lambda h(Q) + (1-\lambda)h(Q') - \frac{\alpha\lambda(1-\lambda)}{2}\|Q - Q'\|^2$$

Seriously, how do we pick H?

- We will consider functions H of the form
$$H(P) = \mathbb{E}_S\left[h\left(P_{|S}\right)\right],$$
where
  - $P_{|S}$ is the c… P
  - h is a conv…
    $\forall Q, Q'$ …
  - h is α-strongly convex wrt some norm $\|\cdot\|^2$:
    $\forall Q, Q' \in \mathrm{Dist}(\mathcal{W}), \lambda \in [0,1]$:
$$h(\lambda Q + (1-\lambda)Q') \leq \lambda h(Q) + (1-\lambda)h(Q') - \frac{\alpha\lambda(1-\lambda)}{2}\|Q - Q'\|^2$$

Terminology:
H: "dependence measure"
h: "conditional dependence measure"

Main result

Theorem

Suppose that H satisfies the conditions above for $\alpha > 0$ and $\|\cdot\|^2$.
Then, for any learning algorithm $\mathcal{A}$,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]| \leq \sqrt{\frac{2H(P_n)\mathbb{E}[\|\ell(\cdot, Z')\|_*^2]}{\alpha n}}$$

Dual norm: $\|\ell(\cdot, Z)\|_* = \sup_{Q-Q' : \|Q-Q'\|=1} \langle Q - Q', \ell(\cdot, Z)\rangle$

Basic examples

Relative entropy

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] \leq \sqrt{\frac{4\mathcal{D}_{\mathrm{KL}}\left(P_{W_n,S_n} | P_{W_n} \otimes P_{S_n}\right)\mathbb{E}\left[\|\bar{\ell}(\cdot, Z')\|_\infty^2\right]}{n}}$$

p-norm with $p \in (1,2]$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] \leq \sqrt{\frac{4\mathbb{E}\left[\|P_{W_n|S_n} - P_{W_n}\|_p^2\right]\mathbb{E}\left[\|\bar{\ell}(\cdot, Z')\|_q^2\right]}{(p-1)n}}$$

p-norm with $p > 2$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] \leq \frac{2p\mathbb{E}\left[\|P_{W_n|S_n} - P_{W_n}\|_p^p\right]\mathbb{E}\left[\|\bar{\ell}(\cdot, Z')\|_q^q\right]}{(p-1)n^{1/p}}$$

Basic examples

Relative entropy

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)] \leq \sqrt{\frac{4\mathcal{D}_{\mathrm{KL}}\left(P_{W_n,S_n} | P_{W_n} \otimes P_{S_n}\right)\mathbb{E}\left[\|\bar{\ell}(\cdot,Z')\|_\infty^2\right]}{n}}$$

p-norm with $p \in (1,2]$

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)] \leq \sqrt{\frac{4\mathbb{E}\left[\|P_{W_n|S_n} - P_{W_n}\|_p^2\right]\mathbb{E}\left[\|\bar{\ell}(\cdot,Z')\|_q^2\right]}{(p-1)n}}$$

These norms remains meaningful for unbounded/heavy tailed losses

p-norm with $p > 2$

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)] \leq \frac{2p\mathbb{E}\left[\|P_{W_n|S_n} - P_{W_n}\|_p^p\right]\mathbb{E}\left[\|\bar{\ell}(\cdot,Z')\|_q^q\right]}{(p-1)n^{1/p}}$$

Basic examples

Relative entropy

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] \le \sqrt{\frac{4\mathcal{D}_{\mathrm{KL}}\left(P_{W_n,S_n} | P_{W_n} \otimes P_{S_n}\right)\mathbb{E}\left[\|\bar{\ell}(\cdot, Z')\|_\infty^2\right]}{n}}$$

p-norm with $p \in (1,2]$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] \le \sqrt{\frac{4\mathbb{E}\left[\|P_{W_n|S_n} - P_{W_n}\|_p^2\right]\mathbb{E}\left[\|\bar{\ell}(\cdot, Z')\|_q^2\right]}{(p-1)n}}$$

p-norm with $p > 2$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] \le \frac{2p\mathbb{E}\left[\|P_{W_n|S_n} - P_{W_n}\|_p^p\right]\mathbb{E}\left[\|\bar{\ell}(\cdot, Z')\|_q^q\right]}{(p-1)n^{1/p}}$$

☹ ☹ ☹ All of these are potentially unbounded / meaningless ☹ ☹ ☹

These norms remains meaningful for unbounded/heavy tailed losses

Outline

- Supervised learning crash course
- Beyond "information theoretic" generalization
- Generalization bounds via convex analysis
- Classic examples: relative entropy, χ^2, p-norm…
- New (and cool?) example: smoothed relative entropy
- Some words about the proof

The smoothed relative entropy

- Let $\mathcal{W} = \mathbb{R}^d$ and define the Gaussian smoothing operator for $\sigma > 0$ on distributions Q over $\mathcal{W}$ as
$$G_\sigma Q = \mathrm{Law}(W + \sigma\xi) \qquad (W \sim Q,\ \xi \sim \mathcal{N}(0, I))$$

The smoothed relative entropy

- Let $\mathcal{W} = \mathbb{R}^d$ and define the Gaussian smoothing operator for $\sigma > 0$ on distributions Q over $\mathcal{W}$ as
$$G_\sigma Q = \mathrm{Law}(W + \sigma\xi) \qquad (W \sim Q,\ \xi \sim \mathcal{N}(0, I))$$
- Define the smoothed relative entropy as
$$\mathcal{D}_\sigma(Q|Q') = \mathcal{D}_{\mathrm{KL}}(G_\sigma Q|G_\sigma Q')$$
and the smoothed total variation distance as
$$\|Q - Q'\|_\sigma = \|G_\sigma Q - G_\sigma Q'\|_{\mathrm{TV}}$$

Smoothing is cool

$$\frac{1}{2}\|Q - Q'\|_{\sigma}^{2} \leq \mathcal{D}_{\sigma}(Q|Q') \leq \frac{1}{2\sigma^{2}}\mathbb{W}_{2}^{2}(Q, Q')$$

Smoothing is cool

$$\frac{1}{2}\|Q - Q'\|_\sigma^2 \leq \mathcal{D}_\sigma(Q|Q') \leq \frac{1}{2\sigma^2}\mathbb{W}_2^2(Q, Q')$$

Theorem

For any learning algorithm $\mathcal{A}$,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]| \leq \sqrt{\frac{\frac{1}{\sigma^2}\mathbb{E}\left[\mathbb{W}_2^2\left(P_{W_n|S_n}, P_{W_n}\right)\right]\mathbb{E}\left[\|\ell(\cdot, Z')\|_{\sigma,*}^2\right]}{n}}$$

Smoothing is cool

$$\frac{1}{2}\|Q - Q'\|_\sigma^2 \leq \mathcal{D}_\sigma(Q|Q') \leq \frac{1}{2\sigma^2}\mathbb{W}_2^2(Q, Q')$$

Theorem

For any learning algorithm $\mathcal{A}$,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]| \leq \sqrt{\frac{\frac{1}{\sigma^2}\mathbb{E}\left[\mathbb{W}_2^2\left(P_{W_n|S_n}, P_{W_n}\right)\right]\mathbb{E}\left[\|\ell(\cdot, Z')\|_{\sigma,*}^2\right]}{n}}$$

When is this small??

The dual norm $\|\cdot\|_{\sigma,*}$

Lemma

Suppose that f is infinitely smooth in the sense that all for all k, all of its partial derivatives of order k are bounded as $|D^k f(w)| \leq \beta_k$.

Then, $\|f\|_{\sigma,*} \leq \sum_{k=0}^{\infty} \left(\sigma\sqrt{d}\right)^k \beta_k$.

The dual norm $\|\cdot\|_{\sigma,*}$

Lemma

Suppose that f is infinitely smooth in the sense that all for all k, all of its partial derivatives of order k are bounded as $\left|D^k f(w)\right| \leq \beta_k$.

Then, $\|f\|_{\sigma,*} \leq \sum_{k=0}^{\infty} \left(\sigma\sqrt{d}\right)^k \beta_k$.

Theorem

Suppose that $\ell(\cdot, z)$ is infinitely smooth with $\beta_k \leq \beta$ $(\forall k)$. Then,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]| \leq \sqrt{\frac{8\beta^2 d\mathbb{E}\left[\mathbb{W}_2^2\left(P_{W_n|S_n}, P_{W_n}\right)\right]}{n}}$$

The dual norm $\|\cdot\|_{\sigma,*}$

Suppose th … all for all k, all of its partia … $\partial^k f(w)\big| \le \beta_k$.

Generalization error of $\mathcal{O}\left(R\beta\sqrt{d/n}\right)$ when all W's have norm bounded by R!

Theorem

Suppose that $\ell(\cdot, z)$ is infinitely smooth with $\beta_k \le \beta$ $(\forall k)$. Then,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]| \le \sqrt{\frac{8\beta^2 d \mathbb{E}\left[\mathbb{W}_2^2\left(P_{W_n|S_n}, P_{W_n}\right)\right]}{n}}$$

Outline

- Supervised learning crash course
- Beyond “information theoretic” generalization
- Generalization bounds via convex analysis
- Classic examples: relative entropy, χ^2, p-norm…
- New (and cool?) example: smoothed relative entropy
- Some words about the proof

Generalization Bounds via Convex Analysis

Gergely Neu

Universitat Pompeu Fabra Barcelona

joint work with Gábor Lugosi

Funded by ERC StG ScaleR

Generalization Bounds via ~~Convex Analysis~~

Online learning

Gergely Neu

Universitat Pompeu Fabra Barcelona

joint work with Gábor Lugosi

Funded by ERC StG ScaleR

Proof idea:
A reduction to online learning

The generalization game

For each $t = 1,2, \dots, n$, **repeat**

- Online learner picks $\tilde{P}_t = \mathrm{Law}(\widetilde{W}_t, S_n) \in \Delta_n \subset \mathcal{P}(\mathcal{W} \times \mathcal{S})$
- Online learner gains reward $\langle \tilde{P}_t, \bar{\ell}_t \rangle = \mathbb{E}\left[\ell\left(\widetilde{W}_t, Z_t\right) - \ell\left(\widetilde{W}_t, Z'\right)\right]$

Proof idea:
A reduction to online learning

The generalization game

For each $t = 1, 2, \dots, n$, **repeat**

- Online learner picks $\tilde{P}_t = \mathrm{Law}(\widetilde{W}_t, S_n) \in \Delta_n \subset \mathcal{P}(\mathcal{W} \times \mathcal{S})$
- Online learner gains reward $\langle \tilde{P}_t, \bar{\ell}_t \rangle = \mathbb{E}\left[\ell(\widetilde{W}_t, Z_t) - \ell(\widetilde{W}_t, Z')\right]$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] = \frac{1}{n} \sum_{t=1}^{n} \langle P_{W_n, S_n} - \tilde{P}_t, \bar{\ell}_t \rangle + \frac{1}{n} \sum_{t=1}^{n} \langle \tilde{P}_t, \bar{\ell}_t \rangle$$

Proof idea:
A reduction to online learning

The generalization game

For each $t = 1,2, \dots, n$, **repeat**

- Online learner picks $\tilde{P}_t = \mathrm{Law}(\widetilde{W}_t, S_n) \in \Delta_n \subset \mathcal{P}(\mathcal{W} \times \mathcal{S})$
- Online learner gains reward $\langle \tilde{P}_t, \bar{\ell}_t \rangle = \mathbb{E}\left[\ell(\widetilde{W}_t, Z_t) - \ell(\widetilde{W}_t, Z')\right]$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)] = \frac{1}{n}\sum_{t=1}^{n}\langle P_{W_n,S_n} - \tilde{P}_t, \bar{\ell}_t \rangle + \frac{1}{n}\sum_{t=1}^{n}\langle \tilde{P}_t, \bar{\ell}_t \rangle$$

R_T =Regret of online learner

G_T = Total gain of online learner

Proof idea II: “Follow the Regularized Leader”

- We “run” FTRL in the generalization game:

$$\tilde{P}_t = \arg\max_{P \in \Delta_n} \left\{ \eta \left\langle P, \sum_{k=1}^{t-1} \bar{\ell}_t \right\rangle - H(P) \right\}$$

Proof idea II: "Follow the Regularized Leader"

- We "run" FTRL in the generalization game:

$$\tilde{P}_t = \arg\max_{P \in \Delta_n} \left\{ \eta \left\langle P, \sum_{k=1}^{t-1} \overline{\ell}_t \right\rangle - H(P) \right\}$$

- Bound the regret of FTRL using the classic analysis:

$$R_T \leq \frac{H(P_n)}{\eta} + \eta \sum_{t=1}^{n} \mathbb{E}\left[\left\| \overline{\ell}_t(\cdot, Z') \right\|_*^2 \right] \sim \sqrt{n H(P_n) \mathbb{E}\left[\left\| \overline{\ell}_t(\cdot, Z') \right\|_*^2 \right]}$$

Proof idea II: "Follow the Regularized Leader"

- We "run" FTRL in the generalization game:

$$\tilde{P}_t = \arg\max_{P \in \Delta_n} \left\{ \eta \left\langle P, \sum_{k=1}^{t-1} \overline{\ell}_t \right\rangle - H(P) \right\}$$

- Bound the regret of FTRL using the classic analysis:

$$R_T \leq \frac{H(P_n)}{\eta} + \eta \sum_{t=1}^{n} \mathbb{E}\left[\left\| \overline{\ell}_t(\cdot, Z') \right\|_*^2 \right] \sim \sqrt{n H(P_n) \mathbb{E}\left[\left\| \overline{\ell}_t(\cdot, Z') \right\|_*^2 \right]}$$

- Hard part: show the gain of online learner is zero:

$$\left\langle \tilde{P}_t, \overline{\ell}_t \right\rangle = \mathbb{E}\left[\ell\left(\widetilde{W}_t, Z_t\right) - \ell\left(\widetilde{W}_t, Z'\right) \right] = 0$$

  - Ingredients: tricky choice of Δ_n and H and exploiting i.i.d.-ness of data

Proof idea III: Construction of Δ_n

- Define ghost samples $S'_n = \{Z'_1, Z'_2, \ldots, Z'_n\}$
- For all i, define
  - “mixed bag” $S^{(i)} = \{Z_1, Z_2, \ldots, Z_i, Z'_{i+1}, \ldots, Z'_n\}$
  - $W_i = \mathcal{A}\left(S^{(i)}\right)$
  - $P_i = \mathrm{law}(W_i, S_n)$
  - $\Delta_i = \mathrm{conv}(\{P_0, P_1, \ldots, P_i\})$

Proof idea IV: Finishing up

- Two ingredients for showing $\langle \tilde{P}_t, \bar{\ell}_t \rangle = 0$:
  - $\tilde{P}_t \in \Delta_{t-1}$ (by construction of H and $\{\Delta_i\}_i$):
    - $\langle P_{t-1}, \bar{L}_{t-1} \rangle = \langle P_t, \bar{L}_{t-1} \rangle = \cdots = \langle P_n, \bar{L}_{t-1} \rangle$
    - $H(P_{t-1}) \leq H(P_t) \leq \cdots \leq H(P_n)$ (Jensen's inequality)
  - For all $\tilde{P} = \text{law}\left(\widetilde{W}, S_n\right) \in \Delta_{t-1}$, we have
$$\langle \tilde{P}, \bar{\ell}_t \rangle = \mathbb{E}\left[\bar{\ell}\left(\widetilde{W}, Z_t\right)\right] = 0$$
(thanks to the independence of $\widetilde{W}$ and Z_t)

Proof idea IV: Finishing up

- Two ingredients for showing $\langle \tilde{P}_t, \bar{\ell}_t \rangle = 0$:
  - $\tilde{P}_t \in \Delta_{t-1}$ (by construction of H and $\{\Delta_i\}_i$):
    - $\langle P_{t-1}, \bar{L}_{t-1} \rangle = \langle P_t, \bar{L}_{t-1} \rangle = \cdots = \langle P_n, \bar{L}_{t-1} \rangle$
    - $H(P_{t-1}) \leq H(P_t) \leq \cdots \leq H(P_n)$ (Jensen's inequality)
  - For all $\tilde{P} = \mathrm{law}(\widetilde{W}, S_n) \in \Delta_{t-1}$, we have

$$\langle \tilde{P}, \bar{\ell}_t \rangle = \mathbb{E}\left[\bar{\ell}\left(\widetilde{W}, Z_t\right)\right] = 0$$

(thanks to the independence of $\widetilde{W}$ and Z_t)

$\blacksquare$

What did we learn & what next?

- We can go beyond standard “information-theoretic” techniques!
- Tradeoffs around strong convexity:
  - Large α $\rightarrow$ large $H(P_n)$
  - Small $||\ell||_*$ $\rightarrow$ large $H(P_n)$

What did we learn & what next?

- We can go beyond standard “information-theoretic” techniques!
- Tradeoffs around strong convexity:
  - Large α $\rightarrow$ large $H(P_n)$
  - Small $||\ell||_*$ $\rightarrow$ large $H(P_n)$
- Examples:
  - Boring: relative entropy, p-norm…
  - Cool: Smoothed relative entropy
  - What else? Wasserstein, Fisher…?
- High-probability bounds?

Thanks!!

Appendix

Proof idea (for strongly convex H)

- Define potential function

$$\Phi(\eta \overline{L}_n) = \sup_{P \in \Delta_n} \{\eta \langle P, \overline{L}_n \rangle - H(P)\}$$

- For all $i \in [n]$, define $\overline{L}_i(w, s) = \frac{1}{n} \sum_{k=1}^{i} \overline{\ell}(w, z_k)$
- Decompose potential: $\Phi(\eta \overline{L}_n) = \sum_{i=1}^{n} \left(\Phi(\eta \overline{L}_i) - \Phi(\eta \overline{L}_{i-1})\right)$

Proof idea (for strongly convex H)

- Define potential function
$$\Phi(\eta \overline{L}_n) = \sup_{P \in \Delta_n} \{\eta \langle P, \overline{L}_n \rangle - H(P)\}$$
- For all $i \in [n]$, define $\overline{L}_i(w, s) = \frac{1}{n} \sum_{k=1}^{i} \overline{\ell}\,(w, z_k)$
- Decompose potential: $\Phi(\eta \overline{L}_n) = \sum_{i=1}^{n} \left(\Phi(\eta \overline{L}_i) - \Phi(\eta \overline{L}_{i-1})\right)$
- Use the convexity + smoothness of $\Phi = H^*$:
$$\Phi(\eta \overline{L}_i) \leq \Phi(\eta \overline{L}_{i-1}) + \langle \nabla \Phi(\eta \overline{L}_{i-1}), \eta \overline{L}_i - \eta \overline{L}_{i-1} \rangle + \frac{\left\| \eta \overline{L}_i - \eta \overline{L}_{i-1} \right\|_*^2}{2\alpha}$$

Proof idea (for strongly convex H)

- Define potential function
$$\Phi(\eta\overline{L}_n) = \sup_{P\in\Delta_n}\{\eta\langle P,\overline{L}_n\rangle - H(P)\}$$
- For all $i \in [n]$, define $\overline{L}_i(w,s) = \frac{1}{n}\sum_{k=1}^{i}\overline{\ell}(w,z_k)$
- Decompose potential: $\Phi(\eta\overline{L}_n) = \sum_{i=1}^{n}\left(\Phi(\eta\overline{L}_i) - \Phi(\eta\overline{L}_{i-1})\right)$
- Use the convexity + smoothness of $\Phi = H^*$:

$$\Phi(\eta\overline{L}_i) \le \Phi(\eta\overline{L}_{i-1}) + \langle\nabla\Phi(\eta\overline{L}_{i-1}), \eta\overline{L}_i - \eta\overline{L}_{i-1}\rangle + \frac{\left\|\eta\overline{L}_i - \eta\overline{L}_{i-1}\right\|_*^2}{2\alpha}$$

Proof idea (for strongly convex H)

- Define potential function
$$\Phi\left(\eta\overline{L}_n\right) = \sup_{P\in\Delta_n}\left\{\eta\left\langle P, \overline{L}_n\right\rangle - H(P)\right\}$$
- For all $i \in [n]$, define $\overline{L}_i(w,s) = \frac{1}{n}\sum_{k=1}^{i} \overline{\ell}\,(w, z_k)$
- Decompose potential: $\Phi\left(\eta\overline{L}_n\right) = \sum_{i=1}^{n}\left(\Phi\left(\eta\overline{L}_i\right) - \Phi\left(\eta\overline{L}_{i-1}\right)\right)$
- Use the convexity + smoothness of $\Phi = H^*$:

$$\Phi\left(\eta\overline{L}_i\right) \le \Phi\left(\eta\overline{L}_{i-1}\right) + \left\langle\nabla\Phi\left(\eta\overline{L}_{i-1}\right), \eta\overline{L}_i - \eta\overline{L}_{i-1}\right\rangle + \frac{\left\|\eta\overline{L}_i - \eta\overline{L}_{i-1}\right\|_*^2}{2\alpha}$$
$$= \frac{\eta^2\left\|\overline{\ell}\right\|_*^2}{2\alpha n^2}$$

Proof idea (for strongly convex H)

- Define potential function

$$\Phi(\eta\overline{L}_n) = \sup_{P\in\Delta_n} \{\eta\langle P, \overline{L}_n\rangle - H(P)\}$$

- For all $i \in [n]$, define $\overline{L}_i(w,s) = \frac{1}{n}\sum_{k=1}^{i} \overline{\ell}\,(w, z_k)$
- Decompose potential: $\Phi(\eta\overline{L}_n) = \sum_{i=1}^{n}\left(\Phi(\eta\overline{L}_i) - \Phi(\eta\overline{L}_{i-1})\right)$
- Use the convexity + smoothness of $\Phi = H^*$:

$$\Phi(\eta\overline{L}_i) \leq \Phi(\eta\overline{L}_{i-1}) + \underbrace{\langle\nabla\Phi(\eta\overline{L}_{i-1}), \eta\overline{L}_i - \eta\overline{L}_{i-1}\rangle}_{=0} + \frac{\left\|\eta\overline{L}_i - \eta\overline{L}_{i-1}\right\|_*^2}{2\alpha}$$

$$= \frac{\eta^2\left\|\overline{\ell}\right\|_*^2}{2\alpha n^2}$$

Construction of Δ_n

- Define ghost samples $S_n' = \{Z_1', Z_2', \dots, Z_n'\}$
- For all i, define
  - "mixed bag" $S^{(i)} = \{Z_1, Z_2, \dots, Z_i, Z_{i+1}', \dots, Z_n'\}$
  - $W_i = \mathcal{A}\left(S^{(i)}\right)$
  - $P_i = \mathrm{law}(W_i, S_n)$
  - $\Delta_i = \mathrm{conv}(\{P_0, P_1, \dots, P_i\})$

Finishing up

- Three ingredients for showing $\langle \nabla\Phi(\eta\bar{L}_{i-1}), \eta\bar{L}_i - \eta\bar{L}_{i-1}\rangle = 0$:
  - $\nabla\Phi(\eta\bar{L}_{i-1}) = \mathrm{argmax}_{P\in\Delta_n}\{\eta\langle P, \bar{L}_{i-1}\rangle - H(P)\}$ (Danskin's theorem)
  - Maximizer is in Δ_{i-1} (by construction of H and $\{\Delta_i\}_i$):
    - $\langle P_{i-1}, \bar{L}_{i-1}\rangle = \langle P_i, \bar{L}_{i-1}\rangle = \cdots = \langle P_n, \bar{L}_{i-1}\rangle$
    - $H(P_{i-1}) \leq H(P_i) \leq \cdots \leq H(P_n)$
  - For all $\tilde{P} = \mathrm{law}(\widetilde{W}, S_n) \in \Delta_{i-1}$, we have

$$\langle \tilde{P}, \eta\bar{L}_i - \eta\bar{L}_{i-1}\rangle = \frac{\eta}{n}\mathbb{E}[\bar{\ell}(\widetilde{W}, Z_i)] = 0$$

(thanks to the independence of $\widetilde{W}$ and Z_i)

Finishing up

- Three ingredients for showing $\langle \nabla\Phi(\eta\overline{L}_{i-1}), \eta\overline{L}_i - \eta\overline{L}_{i-1} \rangle = 0$:
  - $\nabla\Phi(\eta\overline{L}_{i-1}) = \mathrm{argmax}_{P\in\Delta_n}\{\eta\langle P, \overline{L}_{i-1}\rangle - H(P)\}$ (Danskin's theorem)
  - Maximizer is in Δ_{i-1} (by construction of H and $\{\Delta_i\}_i$):
    - $\langle P_{i-1}, \overline{L}_{i-1}\rangle = \langle P_i, \overline{L}_{i-1}\rangle = \cdots = \langle P_n, \overline{L}_{i-1}\rangle$
    - $H(P_{i-1}) \le H(P_i) \le \cdots \le H(P_n)$
  - For all $\tilde{P} = \mathrm{law}(\widetilde{W}, S_n) \in \Delta_{i-1}$, we have

$$\langle \tilde{P}, \eta\overline{L}_i - \eta\overline{L}_{i-1}\rangle = \frac{\eta}{n}\mathbb{E}[\overline{\ell}(\widetilde{W}, Z_i)] = 0$$

(thanks to the independence of $\widetilde{W}$ and Z_i)

■

Strong convexity of $\mathcal{D}_\sigma$

Lemma

The function $h(Q) = \mathcal{D}_\sigma\left(Q \middle| P_{W_n}\right)$ is 1-strongly convex with respect to the smoothed total variation distance.

Proof steps:

- The Bregman divergence of h is $\mathcal{B}_h(Q|Q') = \mathcal{D}_\sigma(Q|Q')$
- Pinsker's inequality:

$$\mathcal{D}_\sigma(Q|Q') = \mathcal{D}_{\mathrm{KL}}(G_\sigma Q|G_\sigma Q') \geq \frac{1}{2}\|G_\sigma Q - G_\sigma Q'\|_{\mathrm{TV}}^2 = \frac{1}{2}\|Q - Q'\|_\sigma^2$$

Boundedness of $\mathcal{D}_\sigma$

Lemma

The smoothed relative entropy is upper-bounded by the squared Wasserstein-2 distance: $\mathcal{D}_\sigma(Q|Q') \leq \frac{1}{2\sigma^2}\mathbb{W}_2^2(Q,Q')$

Proof steps:

- Let π be the coupling of Q and Q' that achieves the infimum in the def. of $\mathbb{W}_2$
- $\mathcal{D}_\sigma(Q|Q') = \mathcal{D}_{\mathrm{KL}}\left(\int_{\mathcal{W}} \mathcal{N}(w,\sigma^2 I)\mathrm{d}\pi(w,w') \,\middle|\, \int_{\mathcal{W}} \mathcal{N}(w',\sigma^2 I)\mathrm{d}\pi(w,w')\right)$

$$\leq \int_{\mathcal{W}} \mathcal{D}_{\mathrm{KL}}(\mathcal{N}(w,\sigma^2 I)|\mathcal{N}(w',\sigma^2 I))\,\mathrm{d}\pi(w,w') = \int_{\mathcal{W}} \frac{1}{2\sigma^2}\|w-w'\|^2\,\mathrm{d}\pi(w,w')$$

Jensen's inequality + joint convexity of $\mathcal{D}_{\mathrm{KL}}$